The Econometrics of Unobservables

– Latent Variable and Measurement Error Models and Their Applications in Empirical Industrial Organization and Labor Economics

Yingyao Hu

The Johns Hopkins University
Department of Economics
yhu@jhu.edu

click here for the latest version [1]

October 13, 2021

[1]©2017, 2021. This manuscript may be printed and reproduced for individual or instructional use, but may not be printed for commercial purposes. Comments are welcome.

to Wei, my best friend forever

Contents

Preface

This manuscript is designed for an advanced micro-econometrics course for graduate students. For empirical researchers, it provides a tool kit to tackle latent variables, such as unobserved heterogeneity, belief, effort, ability, and misreporting errors, in applied microeconomics, especially empirical industrial organization and labor economics. It focuses on nonparametric identification and parametric or semiparametric estimation, and presents specific empirical applications.

The manuscript requires basic knowledge on regression analysis and nonlinear models. I refer to existing books and lecture notes for preparation:

- Bruce Hansen's graduate-level econometrics book ↗
- Jeffrey Wooldridge's graduate-level econometrics textbook:
  Econometric Analysis of Cross Section and Panel Data ↗
- William Greene's graduate-level econometrics textbook:
  Econometric Analysis ↗

I usually start with the following topics:

- Discrete Choice ↗
- Dynamic Discrete Choice ↗

In addition, The presentation slides ↗ of this manuscript are also available.

I plan to keep updating this manuscript, not necessarily for publication, but for the enjoyment of research. Any comments are highly appreciated. Especially, one should feel free to contact me if she or he wants me to cite or discuss her or his work in this manuscript.

And last but not the least, this manuscript is also written for my three kids. From here, they will find out what daddy was doing while they were skating, playing soccer, taking piano, violin, swimming, and karate lessons, studying at AOPS, Kumon, Spidersmart, ...

Structure of Manuscript

This first half of the manuscript presents flexible nonparametric identification and parametric or semiparametric estimation methods for nonlinear models with latent variables. The key methods are extended from the nonclassical measurement error literature.

The second half provides applications of these methods in structural and reduced-form econometrics and in empirical industrial organization and labor economics. These applications involve errors-in-variables, latent variable, unobserved heterogeneity, unobserved state variable, mixture model, hidden Markov model, dynamic discrete choice, unemployment rates, IPV auction, multiple equilibria in incomplete information games, belief, learning model, fixed effects, panel data model, cognitive and non-cognitive skills, matching, income dynamics.

About the companion website

The website ↗ for this file contains:

- A link to freely downloadable latest version of this manuscript and its companion slides.
- Some relevant papers and miscellaneous materials.

Acknowledgements

I am grateful to Yonghong An, Andrew Ching, Yajing Jiang, Zhongjian Lin, Robert Moffitt, Jian Ni, Katheryn Russ, Yuya Sasaki, Tom Wansbeek, Ruli Xiao, Yi Xin, and Tong Zhou for suggestions and comments. All errors are mine.

Yingyao Hu 胡颖尧

1

Introduction

1.1 Observables and Unobservables in Applied Microeconomics

Researchers in applies microeconomics study behavior of economic agents, such as consumers and firms, from observed information in the data. Researchers in this area usually start with an existing microeconomic theoretical models or an intuitive microeconomic relationship, which contains a set of variables to describe economic agents' behavior. This set of variables is composed of three subsets: the agents' decisions or choices, their information set for their decisions, and outside information from which they form their information set.

In the meanwhile, what an empirical researcher observes in the data can be considered as generally-defined measurements of these three subsets of variables, as shown in Table 1.1. When these measurements perfectly reflect the true values, it means that the researcher observes these variables in the data. When these measurements doesn't contain any useful information, it implies the corresponding variables are not observed by the researcher. In the case where these measurements are associated with the latent variables, it means the researcher observes proxies for these latent variables. In most applications, researchers estimate a model based on what they observe in the data, which may contain dependent or endogenous variables and exogenous variables, and treat those unobserved in the data as shocks or error terms. If these unobservables include agents' choices or covariates in agents' information set, their misinterpretation as exogenous errors in the model is a major source of endogeneity.

In this manuscript, we are interested in those variables which are either agents' choices or in the agents' information set but are unobserved to researchers, in particular, those that can't be left in the error terms. If a complete model can fully explain agents' behavior, the main reason for endogeneity in empirical research is due to these unobservables which we focus on. Therefore, the methods we provide here for unobservables also provide a solution to the endogeneity problem, which is arguably the most important problem in econometrics.

Table 1.1: Unobservables of Interest

Variables in micro models	Researchers' data may contain measurements $M(\cdot)$ which are: 1) perfect $M(x) = x$	2) informative M nondegenerated	3) uninformative $M(x) = 0$
Agents' decisions $D(\Omega)$ $D = (Y_1^*, Y_2^*, Y_3^*)$	Dependent var. $Y_1 = Y_1^*$	Proxy $Y_2 \Leftarrow Y_2^*$	Unobs. choices Y_3^*
Agents' information set $\Omega(I)$ $\Omega = (X_1^*, X_2^*, X_3^*)$	Explanatory var. $X_1 = X_1^*$	Proxy $X_2 \Leftarrow X_2^*$	Unobs. covariates X_3^*
Outside information I $I = (\zeta_1, \zeta_2, \zeta_3)$	Instrument var. $Z_1 = \zeta_1$	Noisy IV $Z_2 \Leftarrow \zeta_2$	Shocks ζ_3
	Observables	Obs. $\Leftarrow$ Unobs.	Unobservables

1.2 Why Identification is Important and Challenging

Under the ideal condition, what researchers observe in the data coincides with all the variables in the model of interest. One may directly estimate the model, structural or reduced-form, using a random sample of the variables in the complete model. In many empirical applications, however, there are important variables describing agents' behavior and information set but unobserved by the researchers, such as belief, ability, mood, and effort.

A simple approach to deal with such lack of data information is to assume that these unobservables are independent of the observables and have a known or partially known distribution. The model of interest may then be specified in a likelihood or moment function, in which the unobservables are integrated out.

A reasonable approach is to use additional data information or additional assumptions to identify and estimate the complete model using observed variables by researchers, which may be a subset of the information set of economic agents. Ideally, these additional assumptions, e.g., conditional independence, can be motivated by the economic model. This task is quite challenging due to the existence of unobservables. The identification of complete models with incomplete data information is interesting and important because it lies at the intersection of economic theory and econometric methodology.

Furthermore, we prefer to establish identification before the model of interest is parameterized, which usually leads to local identification and is inherently subject to misspecification. Nonparametric identification allows researchers to answer the following question: Can the economic relationship be revealed by incomplete data information? In the meanwhile, identification of a nonparametric model becomes much more challenging than parametric identification.

From a practitioner's view, ignorance of identification directly lead to inconsistency of an estimator. In layman's terms, we wouldn't know what an estimator is estimating without a solid identification argument. For example, it is well known that many estimators ignoring measurement errors in explanatory variables lead to inconsistent estimates. In addition, non-identification implies a flat likelihood function, with which iterative algorithms may not converge.

1.3 Latent Variable and Measurement Error Models

Latent variable and measurement error models describe the relationship between unobservables and observables. The goal is to identify the distribution of unobservables and also the distribution of observables conditional on unobservables, which corresponds to the distribution of measurement errors. In general, the parameter of interest is the joint distribution, which can be used to describe the relationship between observables and unobservables in economic models.

Early studies on measurement errors in the econometric literature started with the so-called classical measurement error, where the errors are usually assumed to be independent of the true values, arguably because the measurement error models were borrowed from the relevant statistical literature, where the independence assumption is quite reasonable when the measurement error is caused by using an instrument to measure a certain property of an object. The additivity and independence in the classical measurement error models lead to important and fruitful results. In the econometric literature, the classical measurement error framework is adopted mainly for the parsimony of the measurement error part of the model and for the convenience of using existing results. In empirical macroeconomics and some applied microeconomic research, the classical measurement error framework is usually embedded into linear models, such as factor models, linear dynamic models, and linear panel data models. In microeconometrics, identification and estimation of nonlinear models, such as nonlinear regressions and limited dependent models, with classical measurement errors, had been a difficult problem for many years.

In recent years, econometricians have been leading the studies on the nonclassical measurement error model because of the need of handling measurement errors in economic survey data, where the measurement errors are usually caused by self-reporting behaviors. Such a need exists in most disciplines in social sciences. Instead of measuring certain properties of an object, many economic data are from surveys, where interviewees self-report their information. The classical measurement error assumption is unlikely to hold in these scenarios. Econometricians are, therefore, on the frontier of identification and estimation of the so-called nonclassical measurement errors models, where the errors may be correlated with the latent true values. In particular, the presence of nonclassical measurement errors makes the identification of nonlinear models containing the latent true values extremely difficult, that is, whether the models can be uniquely determined from the joint distribution of observed variables.

Based on conditional independence assumptions, which widely exist in economic theo-

ries, a breakthrough in the measurement error models literature has been the realization that the joint distribution of three observables may uniquely determine the joint distribution of four variables including the three observables and the latent variable. Hu (2008) uses a matrix eigenvalue-eigenvector decomposition to show this pathbreaking result for the case where the latent variable is a general discrete variable. The Hu-Schennach Theorem in Hu and Schennach (2008) nontrivially extends this result to the general continuous case using a unique representation of bounded linear operators. In addition, one of the three observables may contain as few information as a binary indicator. Such an identification result is nonparametric and global and leads to a closed-form estimation procedure in the discrete case. The flexibility of these results greatly extend applications of measurement error models to various areas in empirical economic research. This manuscript follows Hu (2017), organizes the existing technical results in terms of the number of measurements, and shows that these technical results may not only apply to measurement error models, but also many economic models with latent variables. For more reviews of this extensive literature, we refer to Wansbeek and Meijer (2000), Bound et al. (2001b), Fuller (2009), Chen et al. (2011), Carroll et al. (2012), Schennach (2016), and Schennach (2019).

2

Nonparametric Identification with Unobservables

This chapter starts with a general framework, where “a measurement" can be simply an observed variable with an informative support. The measurement error distribution describes how observables and unobservables are associated with each other, and contains the information about a mapping from the distribution of the latent variables to the observed measurements. We organize the technical results by the number of measurements needed for identification. In the first example, there are two measurements, which are mutually independent conditioning on the latent variable. With such limited information, strong restrictions on measurement errors are needed to achieve identification in this 2-measurement model. Nevertheless, there are still well known useful results in this framework, such as Kotlarski's identity.

However, when a 0-1 dichotomous indicator of the latent variable is available together with two measurements, nonparametric identification is feasible under a very flexible specification of the model. Hu (2017) names this a 2.1-measurement model, where he uses 0.1 measurement to refer to a 0-1 binary variable. A major breakthrough in the measurement error literature is that the 2.1-measurement model can be non-parametrically identified under mild restrictions (see Hu (2008) and Hu and Schennach (2008)). Since it allows very flexible specifications, the 2.1-measurement model is widely applicable to microeconomic models with latent variables even beyond many existing applications.

Given that any observed random variable can be manually transformed to a 0-1 binary variable, the results for a 2.1-measurement model can be easily extended to a 3-measurement model. A 3-measurement model is useful because many dynamic models involve multiple measurements of a latent variable. A typical example is the hidden Markov model. Results for the 3-measurement model show the exchangeable roles which each measurement may play. In particular, in many cases, it does not matter which one of the three measurements is called a dependent variable, a proxy, or an instrument.

One may also interpret the identification strategy of the 2.1-measurement model as a nonparametric instrumental approach. In that sense, a nonparametric difference-in-differences version of this strategy may help identify more general dynamic processes with

more measurements. As shown in Hu and Shum (2012) , four measurements or four periods of data are enough to identify a rather general partially observed first-order Markov process. Such an identification result is directly applicable to the nonparametric identification of dynamic models with unobserved state variables.

2.1 Definition of a Measurement

In the measurement error literature, researchers usually use the term "measurement" without a formal definition. Here, we adopt the general definition of measurement in Hu (2017). Such a definition is a helpful concept to organize the literature.

Let X denote an observed random variable and X^* be a latent random variable of interest. We define a measurement of X^* as follows:

Definition 1 *A random variable X with support $\mathcal{X}$ is called **a measurement** of a latent random variable X^* with support $\mathcal{X}^*$ if*

$$card\left(\mathcal{X}\right) \geq card\left(\mathcal{X}^*\right),$$

where $card\left(\mathcal{X}\right)$ stands for the cardinality of set $\mathcal{X}$.

The support condition in Definition 1 implies that there exists an injective function from $\mathcal{X}^*$ into $\mathcal{X}$. When X is continuous, the support condition is not restrictive whether X^* is discrete or continuous. When X is discrete, the support condition implies that the number of possible values of one measurement is larger than or equal to that of the latent variable. In addition, the possible values in $\mathcal{X}^*$ are unknown and usually normalized to be the same as those of one measurement with an equal cardinality of the support.

Definition 1 describes a broadly-defined measurement, and doesn't imply or impose any restrictions on how the observables and unobservables are associated with each other. For example, the measurement X defined here can be independent of the true values X^*. How much information the measurement contains about the true values is described in the restrictions imposed on $f_{X|X^*}$, which are introduced below.

X	X^*	
discrete $\{x_1, x_2, ..., x_L\}$	discrete $\{x_1^*, x_2^*, ..., x_K^*\}$	$L \geq K$
continuous	discrete $\{x_1^*, x_2^*, ..., x_K^*\}$	
continuous	continuous	

2.2 A General Framework

In a random sample, we observe measurement X, while the variable of interest X^* is unobserved. The measurement error is defined as the difference $X - X^*$. We can identify the distribution function f_X of measurement X directly from the sample, but our main interest is to identify the distribution of the latent variable f_{X^*}, together with the measurement

error distribution described by $f_{X|X^*}$. The observed measurement and the latent variable are associated as follows: for all $x \in \mathcal{X}$

$$f_X(x) = \int_{\mathcal{X}^*} f_{X|X^*}(x|x^*) f_{X^*}(x^*) dx^*, \tag{2.1}$$

when X^* is continuous and f_{X^*} is the probability density function of X^*, and for all $x \in \mathcal{X} = \{x_1, x_2, \ldots, x_L\}$

$$f_X(x) = \sum_{x^* \in \mathcal{X}^*} f_{X|X^*}(x|x^*) f_{X^*}(x^*), \tag{2.2}$$

when X^* is discrete with support $\mathcal{X}^* = \{x_1^*, x_2^*, \ldots, x_K^*\}$ and $f_{X^*}(x^*) = Pr(X^* = x^*)$ is the probability mass function of X^* and $f_{X|X^*}(x|x^*) = Pr(X = x|X^* = x^*)$. Definition 1 of measurement requires $L \geq K$. We omit arguments of the functions when it does not cause any confusion. This general framework can be used to describe a wide range of economic relationships between observables and unobservables in the sense that the latent variable X^* can be interpreted as unobserved heterogeneity, fixed effects, random coefficients, or latent types in mixture models, etc.

X	measurement	observables
X^*	latent true variable	unobservables

In many empirical models, the latent true variables may have particular economic meanings.

empirical models	unobservables	observables
measurement error	true earnings	self-reported earnings
consumption function	permanent income	observed income
production function	productivity	output, input
wage function	ability	test scores
learning model	belief	choices, proxy
auction model	unobserved heterogeneity	bids
contract model	effort, type	outcome, state var.
...	...	...

For simplicity, we start with the discrete case and define

$$\begin{aligned} \overrightarrow{p}_X &= [f_X(x_1), f_X(x_2), \ldots, f_X(x_L)]^T \\ \overrightarrow{p}_{X^*} &= [f_{X^*}(x_1^*), f_{X^*}(x_2^*), \ldots, f_{X^*}(x_K^*)]^T \\ M_{X|X^*} &= \left[f_{X|X^*}(x_l|x_k^*)\right]_{l=1,2,\ldots,L;k=1,2,\ldots,K}. \end{aligned} \tag{2.3}$$

The notation M^T stands for the transpose of M. Note that $\overrightarrow{p}_X$, $\overrightarrow{p}_{X^*}$, and $M_{X|X^*}$ contain the same information as distributions f_X, f_{X^*}, and $f_{X|X^*}$, respectively. Equation (2.2) is then equivalent to

$$\overrightarrow{p}_X = M_{X|X^*} \overrightarrow{p}_{X^*}. \tag{2.4}$$

The matrix $M_{X|X^*}$ describes the linear transformation from $\mathbb{R}^K$, a vector space containing $\overrightarrow{p}_{X^*}$, to $\mathbb{R}^L$, a vector space containing $\overrightarrow{p}_X$. Suppose that the measurement error distribution, i.e., $M_{X|X^*}$, is known. The identification of the latent distribution f_{X^*} means that if two possible marginal distributions $\overrightarrow{p}^a_{X^*}$ and $\overrightarrow{p}^b_{X^*}$ are observationally equivalent, i.e.,

$$\overrightarrow{p}_X = M_{X|X^*}\overrightarrow{p}^a_{X^*} = M_{X|X^*}\overrightarrow{p}^b_{X^*}, \tag{2.5}$$

then the two distributions are the same, i.e., $\overrightarrow{p}^a_{X^*} = \overrightarrow{p}^b_{X^*}$. Let $h = \overrightarrow{p}^a_{X^*} - \overrightarrow{p}^b_{X^*}$. Equation (2.5) implies that $M_{X|X^*}h = 0$. The identification of f_{X^*} then requires that $M_{X|X^*}h = 0$ implies $h = 0$ for any $h \in \mathbb{R}^K$, or that matrix $M_{X|X^*}$ has rank K, i.e., $Rank\left(M_{X|X^*}\right) = K$. This is a necessary rank condition for the nonparametric identification of the latent distribution f_{X^*}.

In the continuous case, we need to define the linear operator corresponding to $f_{X|X^*}$, which maps f_{X^*} to f_X. Suppose that we know both f_{X^*} and f_X are bounded and integrable. We define $\mathcal{L}^1_{bnd}(\mathcal{X}^*)$ as the set of bounded and integrable functions defined on $\mathcal{X}^*$, i.e.,[1]

$$\mathcal{L}^1_{bnd}(\mathcal{X}^*) = \left\{h : \int_{\mathcal{X}^*} |h(x^*)|\, dx^* < \infty \text{ and } \sup_{x^* \in \mathcal{X}^*} |h(x^*)| < \infty\right\}. \tag{2.6}$$

The linear operator can be defined as

$$\begin{aligned} L_{X|X^*} &: \mathcal{L}^1_{bnd}(\mathcal{X}^*) \to \mathcal{L}^1_{bnd}(\mathcal{X}) \\ \left(L_{X|X^*}h\right)(x) &= \int_{\mathcal{X}^*} f_{X|X^*}(x|x^*)h(x^*)dx^*. \end{aligned} \tag{2.7}$$

Equation (2.1) is then equivalent to

$$f_X = L_{X|X^*}f_{X^*}. \tag{2.8}$$

Following a similar argument, we can show that a necessary condition for the identification of f_{X^*} in the functional space $\mathcal{L}^1_{bnd}(\mathcal{X}^*)$ is that the linear operator $L_{X|X^*}$ is injective, i.e., $L_{X|X^*}h = 0$ implies $h = 0$ for any $h \in \mathcal{L}^1_{bnd}(\mathcal{X}^*)$. This condition can also be interpreted as completeness of conditional density $f_{X|X^*}$ in $\mathcal{L}^1_{bnd}(\mathcal{X}^*)$. We refer to Hu and Schennach (2008) for detailed discussion on this injectivity condition.

Since both the measurement error distribution $f_{X|X^*}$ and the marginal distribution f_{X^*} are unknown, we have to rely on additional restrictions or additional data information to achieve identification. On the one hand, parametric identification may be feasible if $f_{X|X^*}$ and f_{X^*} belong to parametric families (see Fuller (2009)). On the other hand, we can use additional data information to achieve nonparametric identification. For example, if we observe the joint distribution of X and X^* in a validation sample, we can identify $f_{X|X^*}$ from the validation sample and then identify f_{X^*} in the primary sample (see Chen et al. (2005)). In this paper, we focus on methodologies using additional measurements in a single sample.

[1] We may also define the operator on other functional spaces containing f_{X^*}.

2.3 A 2-measurement Model

Given very limited identification results which one may obtain from equations (2.1)-(2.2), a direct extension is to use more data information, i.e., an additional measurement. Define a 2-measurement model as follows:

Definition 2 *A **2-measurement model** contains two measurements, as in Definition 1, $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$ of the latent variable $X^* \in \mathcal{X}^*$ satisfying*

$$X \perp Z \mid X^*, \tag{2.9}$$

i.e., X and Z are independent conditional on X^*.

The 2-measurement model implies that two measurements X and Z not only have distinctive information on the latent variable X^*, but also are mutually independent conditional on the latent variable. In the case where all the variables X, Z, and X^* are discrete with $\mathcal{Z} = \{z_1, z_2, \ldots, z_J\}$, we define

$$\begin{aligned} M_{X,Z} &= [f_{X,Z}(x_l, z_j)]_{l=1,2,...,L;j=1,2,...,J} \\ M_{Z|X^*} &= \left[f_{Z|X^*}(z_j|x_k^*)\right]_{j=1,2,...,J;k=1,2,...,K} \end{aligned} \tag{2.10}$$

and a diagonal matrix

$$D_{X^*} = diag\left\{f_{X^*}(x_1^*), f_{X^*}(x_2^*), \ldots, f_{X^*}(x_K^*)\right\}, \tag{2.11}$$

where $f_{X^*}(x_i^*) > 0$ for $i = 1, 2, ..., K$ by the definition of the discrete support $\mathcal{X}^*$. Definition 1 implies that $K \leq L$ and $K \leq J$. Equation (2.9) means

$$f_{X,Z}(x, z) = \sum_{x^* \in \mathcal{X}^*} f_{X|X^*}(x|x^*) f_{Z|X^*}(z|x^*) f_{X^*}(x^*), \tag{2.12}$$

which is equivalent to

$$M_{X,Z} = M_{X|X^*} D_{X^*} M_{Z|X^*}^T . \tag{2.13}$$

Without further restrictions to reduce the number of unknowns on the right hand side, point identification of $f_{X|X^*}$, $f_{Z|X^*}$, and f_{X^*} may not be feasible. But one element that can be identified from observed $M_{X,Z}$ is the dimension K of the latent variable X^*, as elucidated in the following Lemma:

Lemma 2.3.1 *In the 2-measurement model in Definition 2 with support $\mathcal{X}^* = \{x_1^*, x_2^*, \ldots, x_K^*\}$, suppose that matrices $M_{X|X^*}$ and $M_{Z|X^*}$ both have rank K. Then $K = \text{rank}\,(M_{X,Z})$.*

Proof. In the 2-measurement model, Definition 1 requires that $K \leq L$ and $K \leq J$. The definition of the discrete support $\mathcal{X}^*$ implies that $f_{X^*}(x_i^*) > 0$ for $i = 1, 2, ..., K$ and D_{X^*} has rank K. Using the rank inequality: for any p-by-m matrix A and m-by-q matrix B, $\text{rank}(A) + \text{rank}(B) - m \leq \text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$, we may first show

$M_{X|X^*} D_{X^*}$ has rank K, then use the inequality again to show the right hand side of Equation (2.13) has rank K. Thus, we have rank $(M_{X,Z}) = K$. ■

Point identification of this model requires further restrictions. For example, if $M_{X|X^*}$ and $M^T_{Z|X^*}$ are lower and upper triangular matrices, respectively, point identification is feasible through the so-called LU decomposition (See Hu and Sasaki (2017) for a generalization of such a result). In general, this is also related to the literature on non-negative matrix factorization, which focuses more on existence and approximation, instead of uniqueness.

In the rest of this section, we discuss a class of 2-measurement model, i.e., a regression model with a mismeasured regressor, where one measurement is the dependent variable, the other measurement is the mismeasured regressor. With two observables, we will focus on regression models with an independent regression error.

2.3.1 Regression with a Misclassified Binary Regressor

Although point identification may not be feasible without further assumptions, we can still have some partial identification results. Consider a linear regression model with a discrete regressor X^* as follows:

$$\begin{aligned} Y &= X^*\beta + \eta \\ Y &\perp X \mid X^* \end{aligned} \tag{2.14}$$

where $X^* \in \{0,1\}$ and $E[\eta|X^*] = 0$. Here the dependent variable Y takes the place of Z as a measurement of X^*.[2] We observe (Y, X) with $X \in \{0,1\}$ in the data as two measurements of the latent X^*. Since Y and X are independent conditional on X^*, the two observed distributions with $x = 0, 1$ are different weighted averages of the same two latent distributions, i.e.,

$$f_{Y|X}(y|x) = f_{Y|X^*}(y|0) f_{X^*|X}(0|x) + f_{Y|X^*}(y|1) f_{X^*|X}(1|x). \tag{2.15}$$

Taking the difference with respect to $x = 0, 1$ leads to

$$\begin{aligned} & |E[Y|X^* = 1] - E[Y|X^* = 0]| \\ \geq\ & |E[Y|X = 1] - E[Y|X = 0]|. \end{aligned} \tag{2.16}$$

That means the observed difference provides a lower bound on the parameter of interest $|\beta|$. This is the so-called attenuation phenomenon, as in Figure 2.1. Such a lower bound is useful for testing the hypothesis $\beta = 0$ without further restrictions on the misclassification probability. More partial identification results can be found in Bollinger (1996) and Molinari (2008).

Furthermore, the model can be point identified under the assumption that the regression error η is independent of the regressor X^*. Chen et al. (2009) consider a nonlinear regression

[2] We follow the routine to use Y to denote a dependent variable instead of Z.

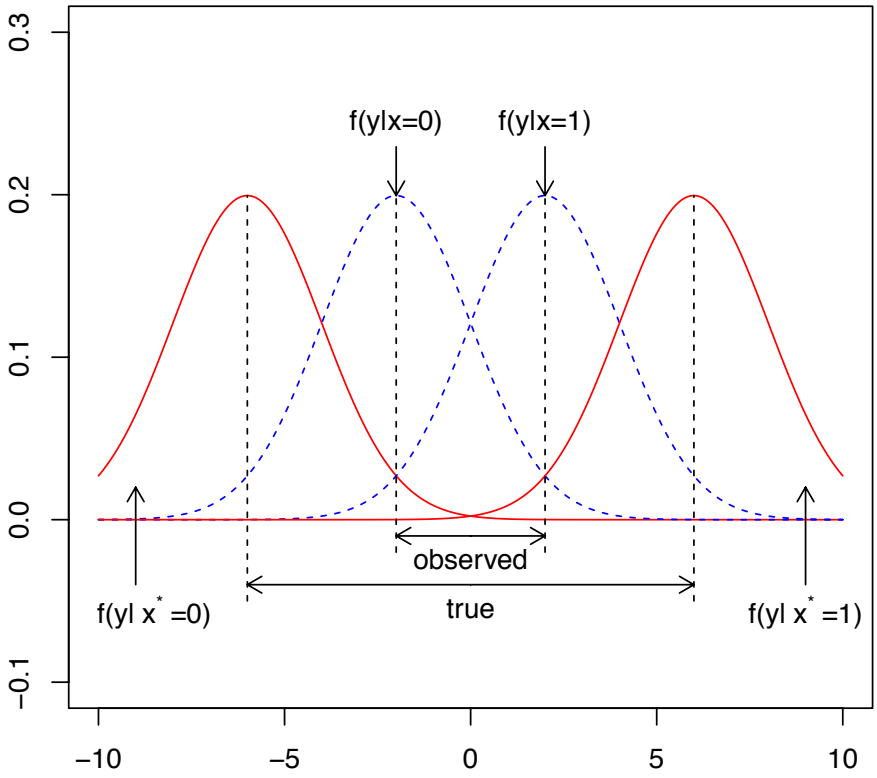


Figure 2.1: Attenuation bias

model with a general discrete X^* as follows:

$$Y = m(X^*) + \eta \tag{2.17}$$

where the regression function m is the unknown of interest. They provide sufficient conditions for identification of m from the joint distribution $f_{Y,X}$ when X^* is independent of η, i.e., $X^* \perp \eta$. In particular, when X^* is 0-1 dichotomous, we have

$$Y = m(0) + [m(1) - m(0)]X^* + \eta. \tag{2.18}$$

Chen et al. (2008b) show that the model can be identified with closed-form expressions. Define

$$\mu_j = E[Y|X = j]$$
$$v_j = E[(Y - \mu_j)^2|X = j]$$
$$s_j = E[(Y - \mu_j)^3|X = j]$$
$$C_1 = \frac{(v_1 + \mu_1^2) - (v_0 + \mu_0^2)}{\mu_1 - \mu_0}$$
$$C_2 = \frac{1}{2}(\mu_1 - \mu_0)^2 + \frac{3}{2}\left(\frac{v_1 - v_0}{\mu_1 - \mu_0}\right)^2 - \frac{s_1 - s_0}{\mu_1 - \mu_0}.$$

Under assumptions that $\mu_1 > \mu_0$ and $f_{X^*|X}(1|0) + f_{X^*|X}(0|1) < 1$, they show the unknown elements of the model can be expressed as closed-form functions of observables as follows:

$$m(0) = \frac{1}{2}C_1 - \sqrt{\frac{1}{2}C_2}$$
$$m(1) = \frac{1}{2}C_1 + \sqrt{\frac{1}{2}C_2}$$
$$f_{X^*|X}(1|0) = \frac{\mu_0 - \frac{1}{2}C_1}{\sqrt{2C_2}} - \frac{1}{2}$$
$$f_{X^*|X}(0|1) = \frac{\frac{1}{2}C_1 - \mu_1}{\sqrt{2C_2}} - \frac{1}{2}$$
$$f_{Y|X^*}(y|j) = \frac{\mu_1 - m(j)}{\mu_1 - \mu_0} f_{Y|X}(y|0) + \frac{m(j) - \mu_0}{\mu_1 - \mu_0} f_{Y|X}(y|1).$$

Such closed-form identification results may be very convenient for empirical researchers.

2.3.2 Regression with a Misclassified Discrete Regressor

Such a point identification result can be extended to a regression model with a general discrete regressor under the assumption that the regression error η is independent of the regressor X^*. Chen et al. (2009) consider a nonlinear regression model with a general discrete X^* as follows:

$$Y = m(X^*) + \eta \tag{2.19}$$

where the regression function m is the unknown of interest. They provide sufficient conditions for identification of m from the joint distribution $f_{Y,X}$ when X^* is independent of η, i.e., $X^* \perp \eta$.

Let X^* have support $\mathcal{X} = \{1, 2, ..., J\}$. They observe a random sample of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where X is a proxy for X^*. The goal is to find restrictions on the latent model $f_{Y|X^*}$ that suffice to nonparametrically identify $f_{Y|X^*}$ and $f_{X|X^*}$ from $f_{Y|X}$.

Assumption 2.3.1 $X \perp \eta | X^*$.

This assumption implies that the measurement error $X - X^*$ is independent of the dependent variable Y conditional on the true value X^*. In other words, we have $f_{Y|X^*,X}(y|x^*, x) = f_{Y|X^*}(y|x^*)$ for all $(x, x^*, y) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y}$. This is equivalent to the classical measurement error property that the outcome Y conditional on both the true X^* and on the measurement error in X, does not depend upon the measurement error.

Assumption 2.3.2 $X^* \perp \eta$.

This assumption implies that the regression error η is independent of the regressor X^* so $f_{Y|X^*}(y|x^*) = f_\eta(y - m(x^*))$. The relationship between the observed density and the latent ones is then

$$f_{Y,X}(y, x) = \sum_{x^*=1}^{J} f_\eta(y - m(x^*)) f_{X,X^*}(x, x^*). \tag{2.20}$$

Assumption 2.3.2 rules out heteroskedasticity or other heterogeneity of the regression error η, but allows its density f_η to be completely unknown and nonparametric. The regression error η is not required to be continuously distributed, but the rank condition discussed below does place a lower bound on the number of points in the support of η. They also show that this assumption can be relaxed in a couple of different ways, e.g., it can be replaced by $E[\exp(it\eta)|X^*, X] = E[\exp(it\eta)]$ for a certain finite set of values of t. For dichotomous (binary) X^*, they show Assumption 2.3.2 can alternatively be weakened to just requiring $E\left(\eta^k | X^*\right) = E\left(\eta^k\right)$ for $k = 2, 3$.

Let ϕ denote a characteristic function (ch.f.). Equation (2.20) is equivalent to

$$\phi_{Y,X=x}(t) = \phi_\eta(t) \sum_{x^*=1}^{J} \exp(itm(x^*)) f_{X,X^*}(x, x^*) \tag{2.21}$$

for all real-valued t, where $\phi_{Y,X=x}(t) = \int \exp(ity) f_{Y,X}(y, x) dy$ and $x \in \mathcal{X}$. Since η may not be symmetric, $\phi_\eta(t) = \int \exp(it\eta) f_\eta(\eta) d\eta$ need not be real-valued. Let $\phi_\eta(t) \equiv |\phi_\eta(t)| \exp(ia(t))$, where

$$|\phi_\eta(t)| \equiv \sqrt{[\mathbf{Re}\{\phi_\eta(t)\}]^2 + [\mathbf{Im}\{\phi_\eta(t)\}]^2}, \quad a(t) \equiv \arccos \frac{\mathbf{Re}\{\phi_\eta(t)\}}{|\phi_\eta(t)|}.$$

We then have for any real-valued scalar t,

$$\phi_{Y,X=x}(t) = |\phi_\eta(t)| \sum_{x^*=1}^{J} \exp(itm(x^*) + ia(t)) f_{X,X^*}(x, x^*). \tag{2.22}$$

Define

$$F_{X,X^*} = \begin{pmatrix} f_{X,X^*}(1,1) & f_{X,X^*}(1,2) & ... & f_{X,X^*}(1,J) \\ f_{X,X^*}(2,1) & f_{X,X^*}(2,2) & ... & f_{X,X^*}(2,J) \\ ... & ... & ... & ... \\ f_{X,X^*}(J,1) & f_{X,X^*}(J,2) & ... & f_{X,X^*}(J,J) \end{pmatrix}.$$

For a real-valued vector $\mathbf{t} = (0, t_2, ..., t_J)$, let $D_{|\phi|}(\mathbf{t}) = Diag\{1, |\phi_\eta(t_2)|, ..., |\phi_\eta(t_J)|\}$,

$$\Phi_{Y,X}(\mathbf{t}) = \begin{pmatrix} f_X(1) & \phi_{Y,X=1}(t_2) & ... & \phi_{Y,X=1}(t_J) \\ f_X(2) & \phi_{Y,X=2}(t_2) & ... & \phi_{Y,X=2}(t_J) \\ ... & ... & ... & ... \\ f_X(J) & \phi_{Y,X=J}(t_2) & ... & \phi_{Y,X=J}(t_J) \end{pmatrix},$$

and take $m_j = m(j)$ for $j = 1, 2, ..., J$, with

$$\Phi_{m,a}(\mathbf{t}) = \begin{pmatrix} 1 & \exp(it_2 m_1 + ia(t_2)) & ... & \exp(it_J m_1 + ia(t_J)) \\ 1 & \exp(it_2 m_2 + ia(t_2)) & ... & \exp(it_J m_2 + ia(t_J)) \\ ... & ... & ... & ... \\ 1 & \exp(it_2 m_J + ia(t_2)) & ... & \exp(it_J m_J + ia(t_J)) \end{pmatrix}.$$

With these matrix notations, for any real-valued vector $\mathbf{t}$, equation (2.22) is equivalent to

$$\Phi_{Y,X}(\mathbf{t}) = F_{X,X^*} \times \Phi_{m,a}(\mathbf{t}) \times D_{|\phi|}(\mathbf{t}). \tag{2.23}$$

Equation (2.23) relates the known parameters $\Phi_{Y,X}(\mathbf{t})$ (which may be interpreted as reduced form parameters of the model) to the unknown structural parameters F_{X,X^*}, $\Phi_{m,a}(\mathbf{t})$, and $D_{|\phi|}(\mathbf{t})$. Equation (2.23) provides a sufficient number of equality constraints to identify the structural parameters given the reduced form parameters, so what is required are sufficient invertibility or rank restrictions to rule out multiple solutions of these equations.

To provide these conditions, consider both the real and imaginary parts of $\Phi_{Y,X}(\mathbf{t})$. Since $D_{|\phi|}(\mathbf{t})$ is real by definition, we have

$$\mathbf{Re}\{\Phi_{Y,X}(\mathbf{t})\} = F_{X,X^*} \times \mathbf{Re}\{\Phi_{m,a}(\mathbf{t})\} \times D_{|\phi|}(\mathbf{t}), \tag{2.24}$$

$$\mathbf{Im}\{\Phi_{Y,X}(\mathbf{t})\} = F_{X,X^*} \times \mathbf{Im}\{\Phi_{m,a}(\mathbf{t})\} \times D_{|\phi|}(\mathbf{t}). \tag{2.25}$$

Since the matrices $\mathbf{Im}\{\Phi_{Y,X}(\mathbf{t})\}$ and $\mathbf{Im}\{\Phi_{m,a}(\mathbf{t})\}$ are not invertible because their first columns are zeros, we replace (2.25) with

$$(\mathbf{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X) = F_{X,X^*} \times (\mathbf{Im}\{\Phi_{m,a}(\mathbf{t})\} + \Upsilon) \times D_{|\phi|}(\mathbf{t}), \tag{2.26}$$

where

$$\Upsilon_X = \begin{pmatrix} f_X(1) & 0 & ... & 0 \\ f_X(2) & 0 & ... & 0 \\ ... & ... & ... & ... \\ f_X(J) & 0 & ... & 0 \end{pmatrix} \text{ and } \Upsilon = \begin{pmatrix} 1 & 0 & ... & 0 \\ 1 & 0 & ... & 0 \\ ... & ... & ... & ... \\ 1 & 0 & ... & 0 \end{pmatrix}.$$

Equation (2.26) holds because $F_{X,X^*} \times \Upsilon = \Upsilon_X$ and $\Upsilon \times D_{|\phi|}(\mathbf{t}) = \Upsilon$. Let $C_{\mathbf{t}} \equiv (\mathbf{Re}\{\Phi_{Y,X}(\mathbf{t})\})^{-1} \times (\mathbf{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X)$.

Assumption 2.3.3 *(rank). There is a real-valued vector* $\mathbf{t} = (0, t_2, ..., t_J)$ *such that (i)* $\boldsymbol{Re}\{\Phi_{Y,X}(\mathbf{t})\}$ *and* $(\boldsymbol{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X)$ *are invertible, and (ii) For any real-valued* $J \times J-$*diagonal matrices* $D_k = Diag\,(0, d_{k,2}, ..., d_{k,J})$*, if* $D_1 + C_{\mathbf{t}} \times D_1 \times C_{\mathbf{t}} + D_2 \times C_{\mathbf{t}} - C_{\mathbf{t}} \times D_2 = 0$*, then* $D_k = 0$ *for* $k = 1, 2$.

Assumption 2.3.3 is analogous to the rank condition for identification in linear models and, in particular, implies identification of the two diagonal matrices

$$D_{\partial \ln|\phi|}(\mathbf{t}) = Diag\left(0, \frac{\partial}{\partial t} \ln|\phi_\eta(t_2)|, ..., \frac{\partial}{\partial t} \ln|\phi_\eta(t_J)|\right),$$

$$D_{\partial a}(\mathbf{t}) = Diag\left(0, \frac{\partial}{\partial t} a(t_2), ..., \frac{\partial}{\partial t} a(t_J)\right).$$

Assumption 2.3.3(ii) is rather complicated, but can be replaced by some simpler sufficient alternatives, which will be described later. Given a candidate value of $\mathbf{t}$, one can test if Assumption 2.3.3 holds for that value, since the assumption is expressed entirely in terms of f_X and the matrix $\Phi_{Y,X}(\mathbf{t})$ which, given a vector $\mathbf{t}$, can be directly estimated from data. It would also be possible to set up a numerical search for sensible candidate values of $\mathbf{t}$ that one might check. For example, letting $Q(\mathbf{t})$ be an estimate of the product of the squared determinants of the matrices in Assumption 2.3.3(i), one could search for values of $\mathbf{t}$ that numerically maximize $Q(\mathbf{t})$. Assumption 2.3.3(i) is then satisfied with high probability if the maximized $Q(\mathbf{t})$ differs significantly from zero.

In the Appendix, Chen et al. (2009) show that

$$\mathbf{Re}\Phi_{Y,X}(\mathbf{t}) \times A_{\mathbf{t}} \times (\mathbf{Re}\Phi_{Y,X}(\mathbf{t}))^{-1} = F_{X|X^*} \times D_m \times \left(F_{X|X^*}\right)^{-1}, \tag{2.27}$$

where $A_{\mathbf{t}}$ on the left-hand side is identified when $D_{\partial \ln|\phi|}(\mathbf{t})$ and $D_{\partial a}(\mathbf{t})$ are identified, $D_m = Diag\,(m(1), ..., m(J))$, and

$$F_{X|X^*} = \begin{pmatrix} f_{X|X^*}(1|1) & f_{X|X^*}(1|2) & ... & f_{X|X^*}(1|J) \\ f_{X|X^*}(2|1) & f_{X|X^*}(2|2) & ... & f_{X|X^*}(2|J) \\ ... & ... & ... & ... \\ f_{X|X^*}(J|1) & f_{X|X^*}(J|2) & ... & f_{X|X^*}(J|J) \end{pmatrix}.$$

Equation (2.27) implies that $f_{X|X^*}(\cdot|x^*)$ and $m(x^*)$ are eigenfunctions and eigenvalues of an identified $J \times J$ matrix on the left. We may then identify $f_{X|X^*}(\cdot|x^*)$ and $m(x^*)$ under the following.

Assumption 2.3.4 *(i)* $m(x^*) < \infty$ *and* $m(x^*) \neq 0$ *for all* $x^* \in \mathcal{X}$*; (ii)* $m(x^*)$ *is strictly increasing in* $x^* \in \mathcal{X}$*.*

Assumption 2.3.4(i) implies that each possible value of X^* is relevant for Y, and 2.3.4(ii) allows us to assign each eigenvalue $m(x^*)$ to its corresponding value x^*. If we only wish to identify the support of the latent factor $m^* = m(X^*)$ and not the regression function $m(\cdot)$ itself, then this monotonicity assumption can be dropped.

Given identification and invertibility of $F_{X|X^*}$, identification of f_{X^*} (the marginal distribution of X^*) immediately follows because f_{X^*} can solved from $f_X = \sum_{X^*} f_{X|X^*} f_{X^*}$ given the invertibility of $F_{X|X^*}$.

Assumption 2.3.4 could be replaced by restrictions on $f_{X|X^*}$ (e.g., by exploiting knowledge about the eigenfunctions rather than eigenvalues to properly assign each $m(x^*)$ to its corresponding value x^*), but Assumption 2.3.4 is more in line with other assumptions, which assume that we have information about the regression model but know very little about the relationship of the unobserved X^* to the proxy X.

Theorem 2.3.1 *Under Assumptions 2.3.1, 2.3.2, 2.3.3 and 2.3.4 in Equation (2.19), the density* $f_{Y,X}$ *uniquely determines* $f_{Y|X^*}$*,* $f_{X|X^*}$*, and* f_{X^*}*.*

Given the model, defined by Assumptions 2.3.1 and 2.3.2, Theorem 2.3.1 shows that Assumptions 2.3.3 and 2.3.4 guarantee that the sample of (Y, X) is informative enough to nonparametrically identify ϕ_η, $m(x^*)$ and f_{X,X^*}, which correspond respectively to the regression error distribution, the regression function, and the joint distribution of the unobserved regressor X^* and the measurement error. This identification is obtained without additional sample information such as an instrumental variable or a secondary sample. Of course, if one has additional covariates such as instruments or repeated measures, they could be exploited along with Theorem 2.3.1. Their results can also be immediately applied if one observes an additional covariate vector W that appears in the regression function, so $Y = m(X^*, W) + \eta$, since their assumptions and results can all be restated as conditioned upon W.

Now consider some simpler sufficient conditions for Assumption 2.3.3(ii) in Theorem 2.3.1. Let $C_{\mathbf{t}}^T$ be the transpose of $C_{\mathbf{t}}$, and "$\circ$" stand for the Hadamard product, i.e., the element-wise product of two matrices.

Assumption 2.3.5 *The real-valued vector* $\mathbf{t} = (0, t_2, ..., t_J)$ *satisfying Assumption 2.3.3(i) also has* $C_{\mathbf{t}} \circ C_{\mathbf{t}}^T + I$ *invertible, and all entries in the first row of the matrix* $C_{\mathbf{t}}$ *nonzero.*

Assumption 2.3.5 implies Assumption 2.3.3(ii), and is in fact stronger than Assumption 2.3.3(ii), since if it holds then one may explicitly solve for $D_{\partial \ln|\phi|}(\mathbf{t})$ and $D_{\partial a}(\mathbf{t})$ in simple closed form. Another alternative to Assumption 2.3.3(ii) is the following

Assumption 2.3.6 *(symmetric rank)* $a(t) = 0$ *for all* t *and, for any real-valued* $J \times J$ *diagonal matrix* $D_1 = Diag(0, d_{1,2}, ..., d_{1,J})$*, if* $D_1 + C_{\mathbf{t}} \times D_1 \times C_{\mathbf{t}} = 0$ *then* $D_1 = 0$*.*

The condition in Assumption 2.3.6 that $a(t) = 0$ for all t is the same as assuming that the distribution of the error term η is symmetric. They call Assumption 2.3.6 the symmetric rank condition because it implies the previous rank condition when η is symmetrically distributed.

Finally, the assumption that the measurement error is independent of the regression error, Assumption 2.3.2, is stronger than necessary. All independence is used for is to obtain (5.3) for some given values of t. More formally, all that is required is that (2.23), and hence (2.25) and (2.26), hold for the vector $\mathbf{t}$ in Assumption 2.3.3. When there are covariates W in the regression model, the requirement becomes that (2.23) hold for the vector $\mathbf{t}$ in Assumption 2.3.3 conditional on W. Therefore, Theorem 2.3.1 holds replacing Assumption 2.3.2 with the following, strictly weaker assumption.

Assumption 2.3.7 *For the known $t = 0, t_2, ..., t_J$ that satisfies Assumption 2.3.3, $\phi_{\eta|X^*=x^*}(t) = \phi_{\eta|X^*=1}(t)$ and $\frac{\partial}{\partial t}\phi_{\eta|X^*=x^*}(t) = \frac{\partial}{\partial t}\phi_{\eta|X^*=1}(t)$ for all $x^* \in \mathcal{X}$.*

This condition permits some correlation of the proxy X with the regression error η, and allows some moments of η to correlate with X^*.

2.3.3 Linear Regression with a Classical Measurement Error

A similar argument also applies to an extended model as follows:

$$\begin{aligned} X &= X^*\beta + \epsilon \\ Z &= X^* + \epsilon'. \end{aligned} \tag{2.28}$$

Suppose $\beta > 0$. A naive OLS estimator obtained by regressing X on Z converges in probability to $\frac{cov(X,Z)}{var(Z)}$, which provides a lower bound on the regression coefficient β. In fact, we have explicit bounds as follows:

$$\frac{cov(X,Z)}{var(Z)} \leq \beta \leq \frac{var(X)}{cov(X,Z)}. \tag{2.29}$$

Furthermore, additional assumptions, such as the joint independence of X^*, ϵ, and ϵ', can lead to point identification of β. Reiersøl (1950) shows that β is point identified when X^* is not normally distributed. To be specific, neither f_{X^*} nor β is identified if and only if X^* is normally distributed and either ϵ' or ϵ can be written as the sum of two independent random variables, one of which is normally distributed.

2.3.4 A Special Case with Closed-Form Solution: Kotlarski's Identity

In the case where all the variables X, Z, and X^* are continuous, a widely-used setup is

$$\begin{aligned} X &= X^* + \epsilon \\ Z &= X^* + \epsilon' \end{aligned} \tag{2.30}$$

where X^*, ϵ, and ϵ' are mutually independent with $E[\epsilon] = 0$. When the error $\epsilon := X - X^*$ is independent of the latent variable X^*, it is called a classical measurement error. This setup is well known because the density of the latent variable X^* can be written as a closed-form function of the observed distribution $f_{X,Z}$. Define $\phi_{X^*}(t) = E\left[e^{itX^*}\right]$ with $i = \sqrt{-1}$ as the characteristic function of X^*. Under the assumption that $\phi_Z(t)$ is absolutely integrable and does not vanish on the real line, we have

$$\begin{aligned} f_{X^*}(x^*) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ix^*t}\phi_{X^*}(t)\,dt \\ \phi_{X^*}(t) &= \exp\left[\int_0^t \frac{iE\left[Xe^{isZ}\right]}{E\left[e^{isZ}\right]}ds\right]. \end{aligned} \tag{2.31}$$

This is the so-called Kotlarski's identity (See Kotlarski (1965) and Rao (1992)). Note that the independence between ϵ and (X^*, ϵ') can be relaxed to a mean independence condition $E[\epsilon|X^*, \epsilon'] = E[\epsilon]$. This identity was first introduced to econometric research by Li and Vuong (1998). Li (2002) first used this result to consistently estimate nonlinear regression models with classical measurement errors. The Kotlarski's identity has been used in many empirical and theoretical studies, including Li et al. (2000) , Krasnokutskaya (2011) , Schennach (2004a), and Evdokimov (2010) .

The intuition of Kotlarski's identity is that the variance of X^* is revealed by the covariance of X and Z, i.e., $var(X^*) = cov(X, Z)$. Therefore, the higher order moments between X and Z can reveal more moments of X^*. If one can pin down all the moments of X^* from the observed moments, the distribution of X^* is then identified under some regularity assumptions.

2.3.5 Nonparametric Regression with a Classical Measurement Error

A more general extension is to consider

$$\begin{aligned} X &= g(X^*) + \epsilon \\ Z &= X^* + \epsilon', \end{aligned} \tag{2.32}$$

where function g is nonparametric and unknown. Schennach and Hu (2013) generalize Reiersøl's result and show that function g and distribution of X^* are nonparametrically identified except for a particular functional form of g or f_{X^*}. The only difference between the model in equation (2.32) and a nonparametric regression model with a classical measurement error is that the regression error ϵ needs to be independent of the regressor X^*.

Schennach and Hu (2013) assume that

Assumption 2.3.8 *The variables X^*, ϵ', ϵ, are mutually independent, $E[\epsilon'] = 0$ and $E[\epsilon] = 0$.*

Assumption 2.3.9 *$E\left[e^{i\xi\epsilon'}\right]$ and $E\left[e^{i\gamma\epsilon}\right]$ do not vanish for any $\xi, \gamma \in \mathbb{R}$, where $i = \sqrt{-1}$.*

Assumption 2.3.10 *$E\left[e^{i\xi X^*}\right]$ and $E\left[e^{i\xi g(X^*)}\right]$ do not vanish for all ξ in a dense subset of $\mathbb{R}$.*

Assumption 2.3.11 *The distribution of X^* admits a uniformly bounded density $f_{X^*}(x^*)$ with respect to the Lebesgue measure.*

Assumption 2.3.12 *The regression function $g(x^*)$ is continuously differentiable over the support of X^*.*

Assumption 2.3.13 *The set $\mathcal{Z} = \{x^* : g'(x^*) = 0\}$ has at most a finite number of elements $x_1^*, \ldots, x_m^*$. If $\mathcal{Z}$ is nonempty, $f_{X^*}(x^*)$ is continuous and nonvanishing in a neighborhood of each x_k^*, $k = 1, \ldots, m$.*

Their main result can then be stated as follows:

Theorem 2.3.2 *(Schennach and Hu (2013)) Let Assumptions 2.3.8-2.3.13 hold.*

1. *If $g(x^*)$ is not of the form*

$$g(x^*) = a + b \ln\left(e^{cx^*} + d\right) \tag{2.33}$$

for some constants $a, b, c, d \in \mathbb{R}$ then $f_{X^*}(x^*)$ and $g(x^*)$ (over the support of $f_{X^*}(x^*)$) in equation 2.32 are identified.

2. *If $g(x^*)$ is of the form (2.33) with[3] $d > 0$, then neither $f_{X^*}(x^*)$ nor $g(x^*)$ in equation 2.32 are identified iff X^* has a density of the form*

$$f_{X^*}(x^*) = A \exp\left(-Be^{Cx^*} + CDx^*\right)\left(e^{Cx^*} + E\right)^{-F} \tag{2.34}$$

with[4] $C \in \mathbb{R}$, $A, B, D, E, F \in [0, \infty[$ and ϵ is decomposable with a type I extreme value factor.[5]

3. *If $g(x^*)$ is linear (i.e. of the form (2.33) with $d = 0$), then neither $f_{X^*}(x^*)$ nor $g(x^*)$ in equation 2.32 are identified iff X^* is normally distributed and either ϵ' or ϵ is decomposable with a normal factor.* [6]

2.3.6 Nonparametric Regression with a Nonclassical Measurement Error

Hu et al. (2021) consider a very general case of nonparametric regressions with nonclassical continuous measurement errors. Let $\mathcal{Y}$, $\mathcal{X}$, and $\mathcal{X}^*$ denote the supports of the distributions

[3] A case where $d < 0$ can be converted into a case with $d > 0$ by permuting the roles of Z and X.

[4] The constants A, B, C, D, E, F depend on a, b, c, d, although this dependence is omitted here for simplicity. Constants yielding a valid density can be found for any a, b, c, d (with $d > 0$).

[5] A type I extreme value distribution has a density of the general form $f(u) = K_1 \exp(K_2 \exp(K_3 u) + K_4 u)$. Here, the constant K_1, K_2, K_3, K_4 are such that $f(u)$ integrates to 1 and has zero mean and may depend on a, b, c, d, although this dependence is omitted here for simplicity.

[6] We say that a random variable r is *decomposable with F factor* if r can be written as the sum of two independent random variables (which may be degenerate), one of which has the distribution F.

of the random variables Y, X, and X^*, respectively. They first assume a boundedness restriction on densities and place some restrictions on the regression error η.

Assumption 2.3.14 *(Restrictions on densities) The joint distribution of the random variable X and X^* admits a density f_{X,X^*} with respect to the Lebesgue measure and the conditional density of the measurement error $f_{X|X^*}$ and marginal density of the true regressor f_{X^*} are bounded by a constant.*

Assumption 2.3.15 *(Restrictions on regression error) We assume that*
(i) (Independence) the regressor error η is independent of the latent true regressor X^*,
(ii) (Zero conditional mean) $E[\eta|X^*] = 0$,
(iii) (Nonvanishing characteristic function) $E[\exp(i\gamma\eta)] \neq 0$ for all $\gamma \in \mathbb{R}$.

Assumption 7.1.1(i) effectively imposes an additively separable structure on the regression error η. This assumption implies that the conditional density $f_{Y|X^*}$ is completely determined by the distribution of the regressor error η and the regression function as follows:

$$f_{Y|X^*}(y|x^*) = f_\eta(y - m_0(X^*)).$$

Assumption 7.1.1(ii) is a standard centering restriction on the model's disturbances.

Let $\mathcal{L}^2(\mathfrak{X}) = \{h : \int_{\mathfrak{X}} |h(x)|^2 dx < \infty\}$. The measurement error satisfies the following:

Assumption 2.3.16 *(Restrictions on Measurement Error) Suppose that*
(i) (Nondifferential error) the observed measurement X is independent of dependent variable Y conditional on the unobserved regressor X^*, i.e., for $\forall(y, x, x^*) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{X}^*$

$$f_{Y|X^*,X}(y|x^*, x) = f_{Y|X^*}(y|x^*).$$

(ii) (Invertibility) For any function $h \in \mathcal{L}^2(\mathcal{X}^*)$, $\int f_{X|X^*}(x|x^*)h(x^*)dx^* = 0$ for all $x \in \mathcal{X}$ implies $h(x^*) = 0$ for almost any $x^* \in \mathcal{X}^*$. On the other hand, for any function $h \in \mathcal{L}^2(\mathcal{X})$, $\int f_{X|X^*}(x|x^*)h(x)dx = 0$ for all $x^* \in \mathcal{X}^*$ implies $h(x) = 0$ for almost any $x \in \mathcal{X}$.
(iii) (Normalization) There exists a known functional G such that $G\left[f_{X|X^*}(\cdot|x^*)\right] = x^*$ for any $x^* \in \mathcal{X}^*$.

Assumption 2.3.16(i) implies that the measurement error is *nondifferential*, that is, $X - X^*$ does not affect the true model, $f_{Y|X^*}$, the distribution of the dependent variable Y conditional on the true value X^*. The observed measurement X thus does not provide any more information about Y than the unobserved regressor X^* already does. Such conditional independence restrictions have been extensively used in the recent years.[7] Note that they allow the measurement error $X - X^*$ to be correlated with the true unobserved regressor X^*, which reflects the presence of potential nonclassical measurement error.

[7]For example, Altonji and Matzkin (2005), Heckman and Vytlacil (2005), and Hoderlein and Mammen (2007).

Assumption 2.3.16(ii) implies that the conditional density $f_{X|X^*}$ is complete in both $\mathcal{X}$ and $\mathcal{X}^*$. This condition is related to the invertibility of the integral operator with kernel $f_{X|X^*}$. Intuitively, assuming completeness of $f_{X|X^*}$ is weaker than assuming independence between X^* and $X - X*$, in the same way the space of invertible matrices is much larger (in terms of dimension) than the space of similarly sized matrices A of the special form $A_{ij} = v_{(j-i)}$ for some vector v.[8] Completeness conditions have recently been employed in the nonparametric IV regression models and nonlinear measurement error models and such conditions are often regarded as high level conditions. Canay et al. (2013a) have shown that the completeness condition is not testable in a nonparametric setting with continuous variables. However, Freyberger (2017) provides a first test for the restricted completeness in a nonparametric instrumental variable model by linking the outcome of the test to consistency of an estimator. Hu et al. (2017) rely on known results regarding the Volterra equation to provide sufficient conditions for completeness conditions for densities with compact support with an accessible interpretation and without specific functional form restrictions.[9]

Assumption 2.3.16(iii) is borrowed from Hu and Schennach (2008), because they also use a spectral decomposition, but with less data information and more restrictions on the regression model. Examples of functional G from Assumption 2.3.16(iii) include the mean, the mode, median, or the τ-th quantile. It implies that a location of the distribution $f_{X|X^*}(\cdot|x^*)$ reveals the true value x^*. This condition also imposes restrictions on the support of x, x^*, and therefore, the measurement error. Those include that zero is in the support of the measurement error and that the cardinality of the support of x can't be smaller than that of x^*.

Finally, they assume the regression function satisfies

Assumption 2.3.17 *(Restrictions on regression function) Suppose that the regression function m_0 is continuous, bounded, and strictly monotonic over support $\mathcal{X}^*$.*

The boundedness constraint can be somewhat restrictive and rules out linear functions when the support $\mathcal{X}^*$ is unbounded. However, if the support of x^* is a bounded interval, Assumption 7.1.2 is a rather mild condition and allows for linear functions.

Their main results is as follows:

Theorem 2.3.3 *Under Assumptions 2.3.14, 2.3.16, 7.1.1, and 7.1.2, given the observed density $f_{Y,X}(y,x)$, the equation*

$$f_{Y,X}(y,x) = \int_{\mathcal{X}^*} f_\eta(y - m_0(X^*)) f_{X|X^*}(x|x^*) f_{X^*}(x^*) dx^*$$

permits a unique solution $(m_0, f_\eta, f_{X|X^*}, f_{X^*}) \equiv \alpha_0$.

[8] This analogy exploits the fact that, in the case of discrete measurement error, the link between the observed distribution of X and the unobserved distribution of X^* can be represented by the multiplication of the vector of unobserved probabilities of the different values of X^* by the mislassification matrix A.

[9] More general discussions of completeness can be found in D'Haultfoeuille (2011), Chen et al. (2013), Andrews (2011), and **?**, Mattner (1993), Newey and Powell (2003) and Blundell et al. (2007b).

The formal proof of this result can be outlined as follows. If one knew the distribution of the model error η, one could recover the joint distribution of $(m_0(X^*), X)$ by a standard deconvolution argument, thanks to Assumptions 7.1.1 and 2.3.16(i). From that distribution, one could then recover m_0 and $f_{X|X^*}$ from the assumed normalization restriction (Assumption 2.3.16(iii)), after exploiting the monotonicity and continuity of m_0 (Assumption 7.1.2).[10] Of course, one does not know, a priori, the distribution of η, but one can, in principle, consider any possible trial distribution to get various possible trial values of m_0 and $f_{X|X^*}$. The key realization is that, whenever the assumed density of η is incorrect, this will be detectable by one of the following occurrences: (i) negative densities for the unobserved variables, (ii) violation of Assumption 2.3.16(ii) (invertibility) or (iii) violation of the boundedness constraint of Assumption 7.1.2.

The Appendix provides another, completely independent, proof of Theorem 2.3.3, which delivers a rather different insight into the identification problem. This alternate proof employs operator techniques similar to those used in Hu and Schennach (2008) and can be summarized as follows. The idea is that the integral Equation 2.3.3 can be cast as a system of operator equivalence relations. Solving this system yields an equivalence between an operator entirely built from observable quantities and a product of unknown operators to be determined. They then show that this factorization can be uniquely determined, because it takes the form of an operator diagonalization identity, i.e., the eigenvalues and eigenfunctions of a known operator yield the different pieces of the product. To ensure uniqueness of this decomposition, they appeal to conditions such as the invertibility and normalization on $f_{X|X^*}$ in Assumptions 2.3.16(ii)&(iii) and the monotonic restriction on m_0 in Assumptions 7.1.2.

Although the monotonicity is a strong restriction, the condition is applicable to many empirical settings. they provide three examples in different areas of economics where monotonicity is a reasonable assumption. The first example is the estimation of the impact of education (X^*) on wages (Y) in which there could be reporting errors in education level. The higher education level the higher wage, which implies a monotonic regression function between the wage offer and the true education level. The second empirical example is in estimating the effect of government subsidies (X^*) on firm R&D investment (Y). The measures of government subsidies may suffer measurement errors because they may be hard to summarize when each firm may receive different types of subsidies. The fact that more government subsidies for firms are likely to increase R&D investments indicates a monotonic relation between them. The third example is the relation between household income (X^*, measured with error) and children health status (Y). Since wealthier families have more resource to promote children health, higher household income tends to be associated with better children health status. In all these three examples, one can use the mode as the functional G in Assumption 2.3.16(iii) because people are more likely to tell the truth for their education level, and household income, and firms are more likely to report the true government subsidies.

[10] In the absence of monotonicity, the measurement error distributions along the X axis for different true values of X^* would mix. As a result, one could not easily identify the measurement error distribution by looking at the distribution of X conditional on the value of $m_0(X^*)$.

The point identification result of Theorem 2.3.3 is not only nonparametric, but also global. This is because they show identification by solving the integral equation directly, in the sense that the identification strategy does not rely on the usual local identification condition that a true parameter value is only distinguishable from those parameters values close to the true one.

Their result is applicable beyond regression settings. In general, one may also consider the observables (Y, X) as two measurements or proxies of the latent variable X^*, an observation which is useful, for instance, in factor models. In many empirical applications, the latent variable may represent unobserved heterogeneity or an individual effect. Their result may then allow for flexible relationships between observables and unobservables to achieve nonparametric identification. In addition, the results can also be straightforwardly extended to the case where an additional error-free covariate vector W appears in the regression function, because the assumptions and results can all be restated as conditioned on W.

Their results prompt the question of whether it would be possible to further extend the identification proof to cover the case where both the dependent variable and the regressor are contaminated by a nonclassical error. However, this would necessitate a one-to-one mapping between the space of bivariate density $f_{YX}(y, x)$ and the much "larger" space of pairs of bivariate functions $(f_{X,X^*}(x, x^*), f_{Y|X^*}(y|x^*))$, which is a highly unlikely possibility.

This is the most general identification result for a 2-measurement model in the continuous case, which has been published so far.

2.4 A 2.1-measurement Model

An arguably surprising result is that we can achieve quite general nonparametric identification of a measurement error model if we observe a little more data information, i.e., an extra binary indicator, than in the 2-measurement model. Define a 2.1-measurement model as follows:

Definition 3 *A **2.1-measurement model** contains two measurements, as in Definition 1, $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$ and a 0-1 dichotomous indicator $Y \in \mathcal{Y} = \{0, 1\}$ of the latent variable $X^* \in \mathcal{X}^*$ satisfying*

$$X \perp Y \perp Z \mid X^*, \tag{2.35}$$

i.e., (X, Y, Z) are jointly independent conditional on X^*.

In this definition, I use "0.1 measurement" to refer to a 0-1 dichotomous indicator of the latent variable. I name it the 2.1-measurement model instead of 3-measurement one in order to emphasize the fact that we only need slightly more data information than the 2-measurement model, given that a binary variable is arguably the least informative measurement, except a constant measurement, of a latent random variable.

2.4.1 The Discrete Case

In the case where X, Z, and X^* are discrete, Definition 1 implies that the supports of observed X and Z are larger than or equal to that of the latent X^*. We start our discussion with the case where the three variables share the same support. We assume

Assumption 2.4.1 *The two measurements X and Z and the latent variable X^* share the same support $\mathcal{X}^* = \{x_1^*, x_2^*, \ldots, x_K^*\}$.*

This condition is not restrictive because the number of possible values in $\mathcal{X}^*$ can be identified, as shown in Lemma 6.1.1, and one can always transform a discrete variable into one with less possible values. We will later discuss that case where supports of measurements X and Z are larger than that of X^*.

The conditional independence in equation (5.23) implies[11]

$$f_{X,Y,Z}(x,y,z) = \sum_{x^* \in \mathcal{X}^*} f_{X|X^*}(x|x^*) f_{Y|X^*}(y|x^*) f_{Z|X^*}(z|x^*) f_{X^*}(x^*). \tag{2.36}$$

For each value of $Y = y$, we define

$$\begin{aligned} M_{X,y,Z} &= [f_{X,Y,Z}(x_i, y, z_j)]_{i=1,2,\ldots,K; j=1,2,\ldots,K} \\ D_{y|X^*} &= \begin{bmatrix} f_{Y|X^*}(y|x_1^*) & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & f_{Y|X^*}(y|x_K^*) \end{bmatrix}. \end{aligned} \tag{2.37}$$

Equation (2.36) is then equivalent to

$$M_{X,y,Z} = M_{X|X^*} D_{y|X^*} D_{X^*} M_{Z|X^*}^T \,. \tag{2.38}$$

Next, we assume

Assumption 2.4.2 *Matrix $M_{X,Z}$ has rank K.*

This assumption is imposed on observed probabilities, and therefore, is directly testable. Equation (2.13), i.e.,

$$M_{X,Z} = M_{X|X^*} D_{X^*} M_{Z|X^*}^T \,. \tag{2.39}$$

then implies $M_{X|X^*}$ and $M_{Z|X^*}$ both have rank K. We then eliminate $D_{X^*} M_{Z|X^*}^T$ to obtain

$$M_{X,y,Z} M_{X,Z}^{-1} = M_{X|X^*} D_{y|X^*} M_{X|X^*}^{-1}. \tag{2.40}$$

This equation implies that the observed matrix on the left hand side has an inherent eigenvalue-eigenvector decomposition, where each column in $M_{X|X^*}$ corresponding to $f_{X|X^*}(\cdot|x_k^*)$

[11] Hui and Walter (1980) first consider the case where the latent variable X^* is binary and show that this identification problem can be reduced to solving a quadratic equation. Mahajan (2006) and Lewbel (2007) also consider this binary case in regression models and treatment effect models. See Section 7.2 for details.

is an eigenvector and the corresponding eigenvalue is $f_{Y|X^*}(y|x_k^*)$. Decompositions with different indexing are observationally equivalent, which can be illustrated as follows:

$$\begin{aligned} M_{X,y,Z}M_{X,Z}^{-1} &= M_{X|X^*}D_{y|X^*}M_{X|X^*}^{-1} \\ &= \begin{pmatrix} f_{X|X^*}(x_1|\clubsuit) & f_{X|X^*}(x_1|\heartsuit) & f_{X|X^*}(x_1|\spadesuit) \\ f_{X|X^*}(x_2|\clubsuit) & f_{X|X^*}(x_2|\heartsuit) & f_{X|X^*}(x_2|\spadesuit) \\ f_{X|X^*}(x_3|\clubsuit) & f_{X|X^*}(x_3|\heartsuit) & f_{X|X^*}(x_3|\spadesuit) \end{pmatrix} \\ &\quad \times \begin{pmatrix} f_{Y|X^*}(y|\clubsuit) & 0 & 0 \\ 0 & f_{Y|X^*}(y|\heartsuit) & 0 \\ 0 & 0 & f_{Y|X^*}(y|\spadesuit) \end{pmatrix} \\ &\quad \times \begin{pmatrix} f_{X|X^*}(x_1|\clubsuit) & f_{X|X^*}(x_1|\heartsuit) & f_{X|X^*}(x_1|\spadesuit) \\ f_{X|X^*}(x_2|\clubsuit) & f_{X|X^*}(x_2|\heartsuit) & f_{X|X^*}(x_2|\spadesuit) \\ f_{X|X^*}(x_3|\clubsuit) & f_{X|X^*}(x_3|\heartsuit) & f_{X|X^*}(x_3|\spadesuit) \end{pmatrix}^{-1} \end{aligned}$$

where [12]

$$\{\clubsuit, \heartsuit, \spadesuit\} \overset{\text{1-to-1}}{\Longleftrightarrow} \{x_1^*, x_2^*, x_3^*\}.$$

In order to achieve a unique decomposition, we require that the eigenvalues are distinctive, and that certain location of distribution $f_{X|X^*}(\cdot|x_k^*)$ reveals the value of x_k^*. We assume

Assumption 2.4.3 *There exists a function $\omega(\cdot)$ such that $E[\omega(Y)|X^* = \overline{x}^*] \neq E[\omega(Y)|X^* = \widetilde{x}^*]$ for any $\overline{x}^* \neq \widetilde{x}^*$ in $\mathcal{X}^*$.*

Assumption 2.4.4 *One of the following conditions holds:*
1) $f_{X|X^*}\left(x_1|x_j^*\right) > f_{X|X^*}\left(x_1|x_{j+1}^*\right)$ for $j = 1, 2, \ldots, K-1$;
2) $f_{X|X^*}(x^*|x^*) > f_{X|X^*}(\widetilde{x}^*|x^*)$ for any $\widetilde{x}^* \neq x^* \in \mathcal{X}^*$;
3) There exists a function $\omega(\cdot)$ such that $E\left[\omega(Y)|X^* = x_j^*\right] > E\left[\omega(Y)|X^* = x_{j+1}^*\right]$ for $j = 1, 2, \ldots, K-1$.

The function $\omega(\cdot)$ may be user-specified, such as $\omega(y) = y$, $\omega(y) = 1(y > y_0)$, or $\omega(y) = \delta(y - y_0)$ for some given y_0.[13] Assumption 2.4.4 2) is consistent with the empirical evidences from the validation study in Table 2.1.

When estimating the model using the eigenvalue-eigenvector decomposition, especially with a continuous Y as later in the paper, it is more convenient to average over Y and use the equation below than directly using Equation (2.36) with a fixed y

$$E[\omega(Y)|X = x, Z = z] f_{X,Z}(x, z) = \sum_{x^* \in \mathcal{X}^*} f_{X|X^*}(x|x^*) E[\omega(Y)|x^*] f_{Z|X^*}(z|x^*) f_{X^*}(x^*). \tag{2.41}$$

[12] Such quaint notations are particularly suitable here because they don't imply any ordering.
[13] When Y is binary, the choice of function $\omega(\cdot)$ does not matter. I state the assumptions in this way so that there is no need to rephrase them later for a general Y.

Table 2.1: Self-reported education x conditional on true education x^* in Kane et al. (1999) (Data source: National Longitudinal Class of 1972 and Transcript data)

$f_{x\|x^*}(x_i\|x_j)$	x^* — true education level		
x — self-reported education	x_1–no college	x_2–some college	x_3–BA$^+$
x_1–no college	0.876	0.111	0.000
x_2–some college	0.112	0.772	0.020
x_3–BA$^+$	0.012	0.117	0.980

If the conditional mean $E[Y|X^*]$ is an object of interest instead of $f_{Y|X^*}$ as in a regression model, we can consider the equation above with $\omega(y) = y$ and relax the conditional independence assumption $f_{Y|X^*,X,Z} = f_{Y|X^*}$ implied in the 2.1-measurement model to a conditional mean independence assumption $E[Y|X^*, X, Z] = E[Y|X^*]$.

We summarize the identification result as follows:

Theorem 2.4.1 *(Hu (2008)) Under assumptions 2.4.1, 2.4.2, 2.4.3, and 2.4.4, the 2.1-measurement model in Definition 3 is non-parametrically identified in the sense that the joint distribution of the three variables (X, Y, Z), i.e., $f_{X,Y,Z}$, uniquely determines the joint distribution of the four variables (X, Y, Z, X^*), i.e., f_{X,Y,Z,X^*}, which satisfies*

$$f_{X,Y,Z,X^*} = f_{X|X^*} f_{Y|X^*} f_{Z|X^*} f_{X^*}. \tag{2.42}$$

A brief proof: The conditional independence in Definition 3 of the 2.1-measurement model implies that Equation (2.38) holds. Assumption 2.4.2 leads to an inherent eigenvalue-eigenvector decomposition in Equation (2.40). Assumption 2.4.3 guarantees that there are K linearly independent eigenvectors. These eigenvectors are conditional distributions, and therefore, are normalized automatically because the column sum of each eigenvector is equal to one. Assumption 2.4.4 pins down the ordering of the eigenvectors or the eigenvalues, i.e., the value of the latent variable corresponding to each eigenvector. Assumption 2.4.4(i) implies that the first row of matrix $M_{X|X^*}$ is decreasing in x_j^* and Assumption 2.4.4(ii) implies that x^* is the mode of distribution $f_{X|X^*}(\cdot|x^*)$. Assumption 2.4.4(i) directly implies an ordering of the eigenvalues. Therefore, each element on the right hand side of Equation (2.40) is uniquely determined by the observed matrix on the left hand side. The eigenvectors reveal the conditional distribution $f_{X|X^*}$ and the identification of other distributions then follows. ■

Theorem 2.4.1, particularly under Assumption 2.4.1, provides an exact identification result with a binary Y in the sense that the number of unknown probabilities is equal to the number of observed probabilities in equation (2.36). Assumption 2.4.1 implies that there are $2K^2 - 1$ observed probabilities in $f_{X,Y,Z}(x, y, z)$ on the left hand side of equation (2.36). On the right hand side, there are $K^2 - K$ unknown probabilities in each of $f_{X|X^*}(x|x^*)$ and $f_{Z|X^*}(z|x^*)$, $K - 1$ in $f_{X^*}(x^*)$, and K in $f_{Y|X^*}(y|x^*)$ when Y is binary, which sum up to $2K^2 - 1$. More importantly, this point identification result is nonparametric, global, and constructive. It is constructive in the sense that an estimator can directly mimic the

identification procedure.

When supports of measurements X and Z are larger than that of X^*, we can still achieve the identification with minor modification of the conditions. Suppose supports $\mathcal{X}$ and $\mathcal{Z}$ are larger than $\mathcal{X}^*$, i.e., $\mathcal{X} = \{x_1, x_2, \ldots, x_L\}$, $\mathcal{Z} = \{z_1, z_2, \ldots, z_J\}$, and $\mathcal{X}^* = \{x_1^*, x_2^*, \ldots, x_K^*\}$ with $L > K$ and $J > K$. By combining some values in the supports of X and Z, we first transform X and Z to $\widetilde{X}$ and $\widetilde{Z}$ so that they share the same support $\mathcal{X}^*$ as X^*. We then identify $f_{\widetilde{X}|X^*}$ and $f_{\widetilde{Z}|X^*}$ by Theorem 2.4.1 with those assumptions imposed on $\left(\widetilde{X}, Y, \widetilde{Z}, X^*\right)$. However, the joint distribution f_{X,Y,Z,X^*} may still be of interest. In order to identify $f_{Z|X^*}$ or $M_{Z|X^*}$, we consider the joint distribution

$$f_{\widetilde{X},Z} = \sum_{x^* \in \mathcal{X}^*} f_{\widetilde{X}|X^*} f_{Z|X^*} f_{X^*}, \tag{2.43}$$

which is equivalent to

$$M_{\widetilde{X},Z} = M_{\widetilde{X}|X^*} D_{X^*} M_{Z|X^*}^T \ . \tag{2.44}$$

Since we have identified $M_{\widetilde{X}|X^*}$ and D_{X^*}, we can identify $M_{Z|X^*}$, i.e., $f_{Z|X^*}$, by inverting $M_{\widetilde{X}|X^*}$. Similar argument holds for identification of $f_{X|X^*}$. This discussion implies that Assumptions 2.4.1 is not necessary. We keep it in Theorem 2.4.1 in order to show minimum data information needed for nonparametric identification of the 2.1-measurement model. We summarize this result as follows:

Corollary 2.4.1 *Suppose that there exists a transformation $\widetilde{X}$ of X, which shares the same support as X^* and has a conditional distribution $f_{\widetilde{X}|X^*}$ satisfying Assumption 2.4.4 as $f_{X|X^*}$. Then, Theorem 2.4.1 holds without Assumption 2.4.1.*

2.4.2 Misclassification versus Finite Mixture

Mixture structures arise with the presence of a latent variable, which could be a variable measured with error or unobserved heterogeneity of different sources such as heterogeneous preferences, unobserved heterogeneity within/across markets, different types of beliefs, and multiple equilibria in games. Both finite mixture and misclassification models can be reformulated into similar mixture structures and are widely used in economic applications such as labor economics, industrial organization, and so forth. For example, (Keane and Wolpin, 1997) consider unobserved type-specific endowments; (Hu et al., 2013a) control for auction-level unobserved heterogeneity; and (Xiao, 2018) controls for the presence of multiple equilibria in games. See (Compiani and Kitamura, 2016) for a review of finite mixture models.

Both literatures of finite mixture and misclassification models recover the unobserved component-specific distributions through joint distribution of observables, but they rely on different conditions. A vast literature studies identification and estimation in the two areas. The finite mixture literature initially focus on identification of the latent distributions from the observed distribution by imposing restrictions on the component distribution. For example, the identification is feasible when the component distributions belong to a

parametric family (Everitt and David, 1981) or is symmetric ((Bordes et al., 2006) and (Hunter et al., 2007)). Arguably, because these restrictions are implausible in empirical applications, the conditional independence assumption was introduced later in the finite mixture model with a multi-covariate observable. Such a setup is equivalent to the long-existing misclassification model with multiple measurements. In that sense, one may either interpret the misclassification model as an example of a finite mixture model, or observe that the finite mixture setup is merging into the misclassification model. More importantly, this connection means that the existing powerful results for misclassification models are also applicable to finite mixture models.

Both literatures share the same prevalent label swapping issue, but they address the issue differently in accordance with their respective interpretation of the latent variable. In particular, since the latent variable in misclassification models usually carries economic implications, additional conditions are imposed to pin down the precise value of the latent variable. In contrast, the unobserved component in finite mixture models does not convey any economic meaning, so precise location of the unobserved component is not necessarily required. Consequently, misclassification models reach global identification while finite mixture models reach local identification.

A problem arises with local identification when researchers attempt to use bootstrap to estimate the standard errors of the estimators. Without an appropriate ordering condition, the estimator would be a local one in the sense that multiple estimators can generate the same values for the chosen criteria function; thus, it is not straightforward which local estimator should be chosen for each bootstrap resampling. The existing literature on finite mixture models has realized the importance and necessity of pinning down the component order when standard error is estimated through resampling. For instance, (Kasahara and Shimotsu, 2009) propose determining this component ordering by using the marginal distribution of the component to uniquely pin down the order. (Hall and Zhou, 2003) also suggest similar treatment. (Bonhomme et al., 2016b) note that the label swapping issue presents a challenge for inference methods based on resampling algorithms such as bootstrap. In line with this literature, we advocate imposing a condition to pin down the order of the latent components by which a global estimator may be obtained, as in misclassification models. To this end, finite mixture models are very similar to misclassification models.

The literature on finite mixture models started with a setup as follows:

$$f_D = \sum_{\tau \in \{1,2,...,T\}} f_{D|\tau} f_\tau. \tag{2.45}$$

where $\tau \in \{1, 2, ..., T\}$ for some finite T and D stands for observed variables in the data. Researchers are interested in the distribution $f_{D|\tau}$ while only f_D is observed.

An approach of finite mixture models considers what restrictions can be imposed on the distributions $f_{D|\tau}$ for a small T, e.g., $T = 2$, so that $f_{D|\tau}$ can be uniquely determined by f_D. For example, one of such restrictions may be that $f_{D|\tau}$ is symmetric.

Since such restrictions may be too restrictive for empirical research, another approach of finite mixture models impose conditional independence restrictions such as $D = (X, Y, Z)$

and

$$f_D = \sum_{\tau} f_{X|\tau} f_{Y|\tau} f_{Z|\tau} f_{\tau}. \tag{2.46}$$

Such a setup is literally equivalent to a misclassification model, where many existing identification results apply.

A general local identification result, without the ordering conditions in Assumption 2.4.4 and the support condition in Definition 1, may be found in Allman et al. (2009). In our 2.1-measurement model, the equality in the rank condition in their Theorem 1 holds. To be specific, Assumption 2.4.3, which guarantees distinctive eigenvalues, holds if and only if the so-called Kruskal rank of their matrix corresponding to the binary Y is equal to 2. The Kruskal ranks of their other two matrices are equal to the regular matrix rank K, and therefore, the total Kruskal rank equals $2K + 2$. In addition, for a general discrete Y, Assumption 2.4.3 implies that the Kruskal rank of their matrix corresponding to Y is at least 2.

We prove the claims above as follows. We may define the matrix corresponding to the variable Y in the same way as in Allman et al. (2009) as follows:

$$M_{Y|X^*}^T = \begin{pmatrix} f_{Y|X^*}(0|x_1^*) & f_{Y|X^*}(1|x_1^*) \\ f_{Y|X^*}(0|x_2^*) & f_{Y|X^*}(1|x_2^*) \\ \cdots & \cdots \\ f_{Y|X^*}(0|x_K^*) & f_{Y|X^*}(1|x_K^*) \end{pmatrix}$$

For a matrix M, the Kruskal rank of M will mean the largest number I such that every set of I rows of M are independent. The Kruskal rank is smaller than or equal to the regular rank of the same matrix. In the case where a matrix M of size K-by-L has rank K, it also has Kruskal rank K. That means the Kruskal rank of $M_{X|X^*}^T$ and $M_{Z|X^*}^T$ is K.

We may then show that Assumption 2.4.3 holds if and only if the Kruskal rank of $M_{Y|X^*}^T$ is equal to 2. Let $\omega(x) = x$. Assumption 2.4.3 becomes $f_{Y|X^*}(1|\widetilde{x}^*) - f_{Y|X^*}(1|\overline{x}^*) \neq 0$ for any $\overline{x}^* \neq \widetilde{x}^*$ in $\mathcal{X}^*$. For any 2 rows of $M_{y|x^*}^T$ corresponding to $\overline{x}^* \neq \widetilde{x}^*$, we consider the following matrix

$$\begin{pmatrix} f_{Y|X^*}(0|\overline{x}^*) & f_{Y|X^*}(1|\overline{x}^*) \\ f_{Y|X^*}(0|\widetilde{x}^*) & f_{Y|X^*}(1|\widetilde{x}^*) \end{pmatrix}$$

The determinant of this matrix is

$$\begin{aligned} & \left(1 - f_{Y|X^*}(1|\overline{x}^*)\right) f_{Y|X^*}(1|\widetilde{x}^*) - f_{Y|X^*}(1|\overline{x}^*) \left(1 - f_{Y|X^*}(1|\widetilde{x}^*)\right) \\ = \; & f_{Y|X^*}(1|\widetilde{x}^*) - f_{Y|X^*}(1|\overline{x}^*) \end{aligned}$$

Therefore, Assumption 2.4.3 implies that any 2 rows of $M_{Y|X^*}^T$ are independent. Since Y is binary, the largest number of independent rows is 2. Therefore, the Kruskal rank of $M_{Y|X^*}^T$ is 2. The reverse argument also holds. If the Kruskal rank of $M_{Y|X^*}^T$ is 2, any two rows of that matrix are independent. Therefore, the determinant of the 2-by-2 matrix formed by these two rows is not equal to zero, which implies Assumption 2.4.3 with $\omega(x) = x$.

For a general discrete Y with support $\{y_1, y_2, \ldots, y_m\}$. We may consider

$$M^T_{Y|X^*} = \begin{pmatrix} f_{Y|X^*}(y_1|x^*_1) & f_{Y|X^*}(y_2|x^*_1) & ... & f_{Y|X^*}(y_m|x^*_1) \\ f_{Y|X^*}(y_1|x^*_2) & f_{Y|X^*}(y_2|x^*_2) & ... & f_{Y|X^*}(y_m|x^*_2) \\ ... & ... & ... & ... \\ f_{Y|X^*}(y_1|x^*_K) & f_{Y|X^*}(y_2|x^*_K) & ... & f_{Y|X^*}(y_m|x^*_K) \end{pmatrix}$$

We can show that the Kruskal rank of $M^T_{Y|X^*}$ is at least 2 if and only if for any $\overline{x}^* \neq \widetilde{x}^*$ there exists a y_j such that $f_{Y|X^*}(y_j|\widetilde{x}^*) - f_{Y|X^*}(y_j|\overline{x}^*) \neq 0$. For any two rows with $\overline{x}^* \neq \widetilde{x}^*$, we consider the following matrix

$$M_2 = \begin{pmatrix} f_{Y|X^*}(y_1|\overline{x}^*) & f_{Y|X^*}(y_2|\overline{x}^*) & ... & f_{Y|X^*}(y_m|\overline{x}^*) \\ f_{Y|X^*}(y_1|\widetilde{x}^*) & f_{Y|X^*}(y_2|\widetilde{x}^*) & ... & f_{Y|X^*}(y_m|\widetilde{x}^*) \end{pmatrix}$$

Let $\mathbf{1} = (1, 1, ..., 1)^T$ and $e_j = (0, ..., 0, 1, 0, ..., 0)^T$, where 1 is at the j-th coordinate. We consider

$$M_2 \times (e_j \quad \mathbf{1}) = \begin{pmatrix} f_{Y|X^*}(y_j|\overline{x}^*) & 1 \\ f_{Y|X^*}(y_j|\widetilde{x}^*) & 1 \end{pmatrix}$$

Therefore, the rank of M_2 equals 2 if $f_{Y|X^*}(y_j|\widetilde{x}^*) - f_{Y|X^*}(y_j|\overline{x}^*) \neq 0$. That means the Kruskal rank of $M^T_{Y|X^*}$ is at least 2.

There reverse argument can also be shown similarly. If the Kruskal rank of $M^T_{Y|X^*}$ is at least 2, the rank of matrix M_2 equals 2. That means there must exist a column, say j, in M_2 such that $f_{Y|X^*}(y_j|\widetilde{x}^*) - f_{Y|X^*}(y_j|\overline{x}^*) \neq 0$.

For a more formal description of these results, see Hu and Xiao (forthcoming).

2.4.3 A Geometric Illustration

Given that a matrix is a linear transformation from one vector space to another, we provide a geometric interpretation of the identification strategy. Consider $K = 3$ and define

$$\begin{aligned} \overrightarrow{p}_{X|x^*_i} &= \left[f_{X|X^*}(x_1|x^*_i), f_{X|X^*}(x_2|x^*_i), f_{X|X^*}(x_3|x^*_i)\right]^T \\ \overrightarrow{p}_{X|z} &= \left[f_{X|Z}(x_1|z), f_{X|Z}(x_2|z), f_{X|Z}(x_3|z)\right]^T. \end{aligned} \tag{2.47}$$

We have for each z

$$\overrightarrow{p}_{X|z} = \sum_{i=1}^{3} w^z_i \left(\overrightarrow{p}_{X|x^*_i}\right) \tag{2.48}$$

with $w^z_i = f_{X^*|Z}(x^*_i|z)$ and $w^z_1 + w^z_2 + w^z_3 = 1$. That means each observed distribution of X conditional on $Z = z$ is a weighted average of $\overrightarrow{p}_{X|x^*_1}$, $\overrightarrow{p}_{X|x^*_2}$, and $\overrightarrow{p}_{X|x^*_3}$. Similarly, if we consider the subsample with $Y = 1$, we have

$$\overrightarrow{p}_{y_1,X|z} = \sum_{i=1}^{3} w^z_i \left(\lambda_i \overrightarrow{p}_{X|x^*_i}\right) \tag{2.49}$$

where $\lambda_i = f_{Y|X^*}(1|x_i^*)$ and

$$\overrightarrow{p}_{y_1,X|z} = \left[f_{Y,X|Z}(1,x_1|z), f_{Y,X|Z}(1,x_2|z), f_{Y,X|Z}(1,x_3|z)\right]^T. \tag{2.50}$$

That means vector $\overrightarrow{p}_{y_1,X|z}$ is a weighted average of $\left(\lambda_i \overrightarrow{p}_{X|x_i^*}\right)$ for $i = 1, 2, 3$, where weights w_i^z are the same as in equation (2.48) from the whole sample. Notice that the direction of basis vectors $\left(\lambda_i \overrightarrow{p}_{X|x_i^*}\right)$ corresponding to the subsample with $Y = 1$ is the same as the direction of basis vectors $\overrightarrow{p}_{X|x_i^*}$ corresponding to the whole sample. The only difference is the length of the basis vectors. Therefore, if we consider a mapping from the vector space spanned by $\overrightarrow{p}_{X|z}$ to one spanned by $\overrightarrow{p}_{y_1,X|z}$, the basis vectors do not vary in direction so that they are called eigenvectors, and the variation in the length of these basis vectors is given by the corresponding eigenvalues, i.e., λ_i. This mapping is in fact $M_{X,y,Z}M_{X,Z}^{-1}$ on the left hand side of equation (2.40). The variation in variable Z guarantees that such a mapping exists. Figure 2.2 illustrates this framework.

2.4.4 The Continuous Case

In the case where X, Z, and X^* are continuous, the identification strategy still work by replacing matrices with integral operators. We state assumptions as follows:

Assumption 2.4.5 *The joint distribution of (X, Y, Z, X^*) admits a bounded density with respect to the product measure of some dominating measure defined on $\mathcal{Y}$ and the Lebesgue measure on $\mathcal{X} \times \mathcal{X}^* \times \mathcal{Z}$. All marginal and conditional densities are also bounded.*

Assumption 2.4.6 *The operators $L_{X|X^*}$ and $L_{Z|X}$ are injective.*[14]

Assumption 2.4.7 *For all $\overline{x}^* \neq \widetilde{x}^*$ in $\mathcal{X}^*$, the set $\left\{y : f_{Y|X^*}(y|\overline{x}^*) \neq f_{Y|X^*}(y|\widetilde{x}^*)\right\}$ has positive probability.*

Assumption 2.4.8 *There exists a known functional M such that $M\left[f_{X|X^*}(\cdot|x^*)\right] = x^*$ for all $x^* \in \mathcal{X}^*$.*

Assumption 2.4.6 is a high-level technical condition. A sufficient condition for the injectivity of $L_{Z|X}$ is that the only function $h(\cdot)$ satisfying $E[h(X)|Z = z] = 0$ for any $z \in \mathcal{Z}$ is $h(\cdot) = 0$ over $\mathcal{X}$. This condition is also equivalent to the completeness of the density $f_{X|Z}$ over certain functional space. Assumption 2.4.7 requires that each possible value of the latent variable X^* affects the distribution of Y. The functional $M[\cdot]$ in Assumption 2.4.8 may be mean, mode, median, or another quantile, which maps a probability distribution to a point on the real line. In particular, the zero mode and zero median assumptions are consistent with the empirical evidences from validation studies in Figures 2.3, 2.4, and 2.5.

We summarize the results as follows:

[14] $L_{Z|X}$ is defined in the same way as $L_{X|X^*}$ in equation (2.7).

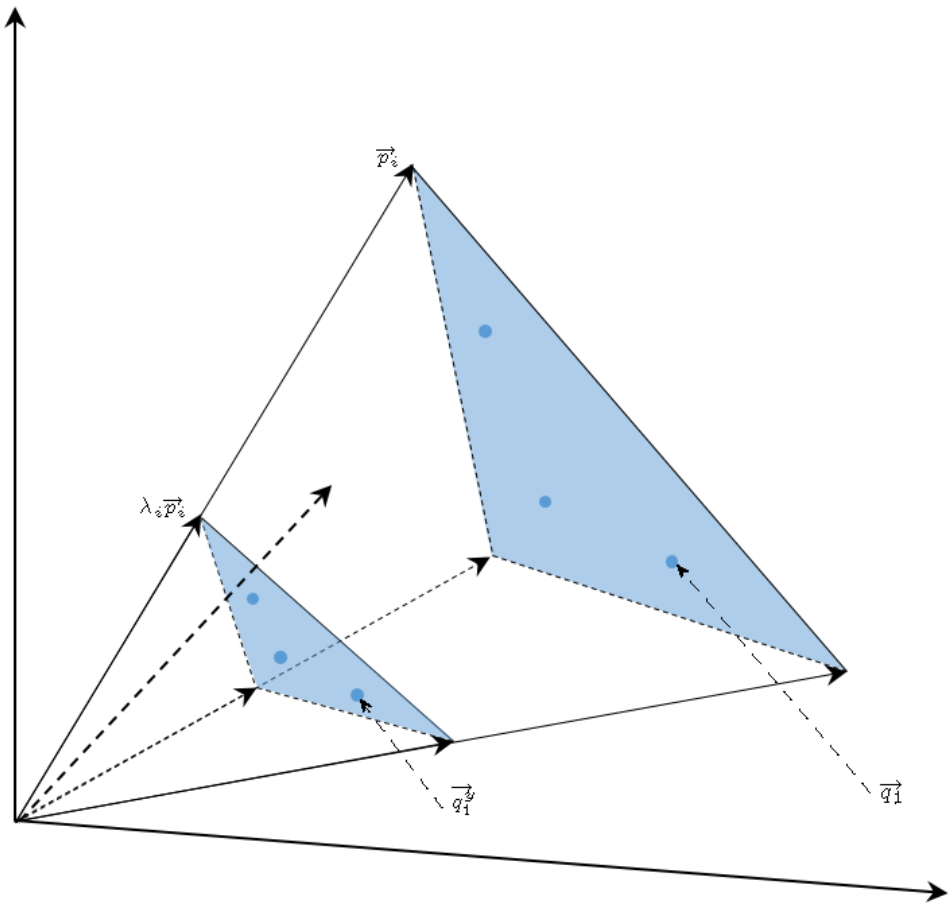


Figure 2.2: Eigenvalue-eigenvector decomposition in the 2.1-measurement model.

Note:
Eigenvalue: $\lambda_i = f_{Y|X^*}(1|x_i^*)$.
Eigenvector: $\overrightarrow{p_i} = \overrightarrow{p}_{X|x_i^*} = \left[f_{X|X^*}(x_1|x_i^*), f_{X|X^*}(x_2|x_i^*), f_{X|X^*}(x_3|x_i^*)\right]^T$.
Observed distribution in the whole sample:
$\overrightarrow{q}_1 = \overrightarrow{p}_{X|z_1} = \left[f_{X|Z}(x_1|z_1), f_{X|Z}(x_2|z_1), f_{X|Z}(x_3|z_1)\right]^T$.
Observed distribution in the subsample with $Y = 1$:
$\overrightarrow{q}_1^y = \overrightarrow{p}_{y_1,X|z_1} = \left[f_{Y,X|Z}(1, x_1|z_1), f_{Y,X|Z}(1, x_2|z_1), f_{Y,X|Z}(1, x_3|z_1)\right]^T$.

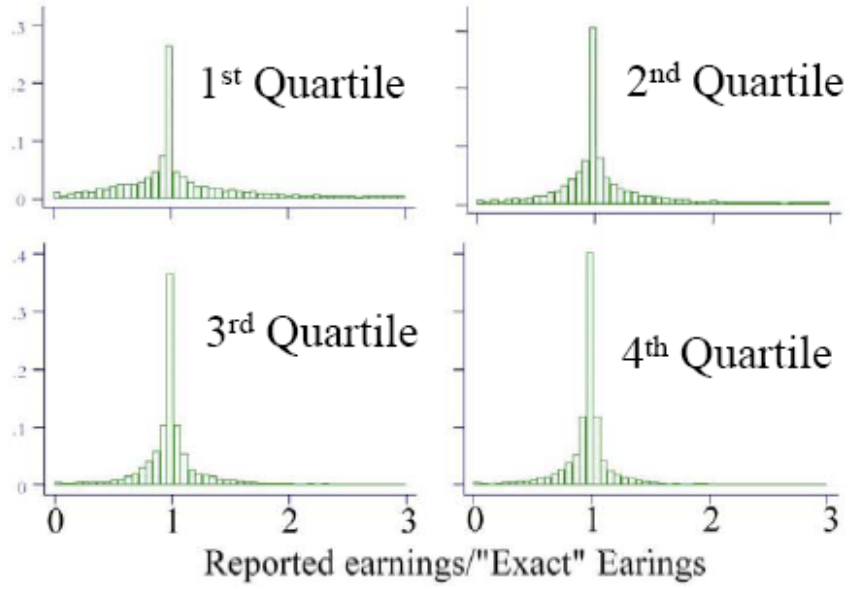


Figure 2.3: Chen et al. (2008a) (page 50): Ratio of self-reported earnings x vs. true earnings x^* by quartiles of true earnings. The link of the paper is: http://cowles.econ.yale.edu/P/cd/d16a/d1644.pdf (Data source: 1978 CPS/SS Exact Match File)

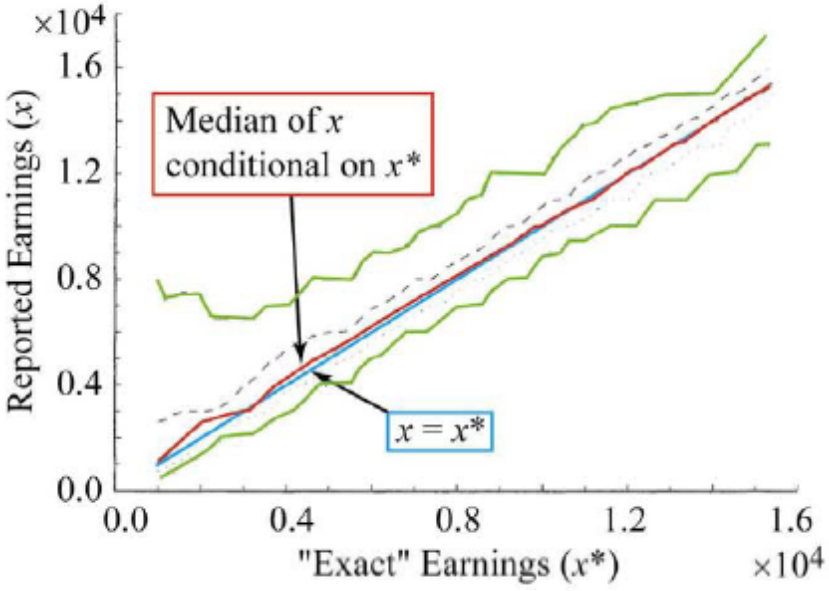


Figure 2.4: Bollinger (1998) (page 591): percentiles of self-reported earnings x given true earnings x^* for males. (Data source: 1978 CPS/SS Exact Match File)

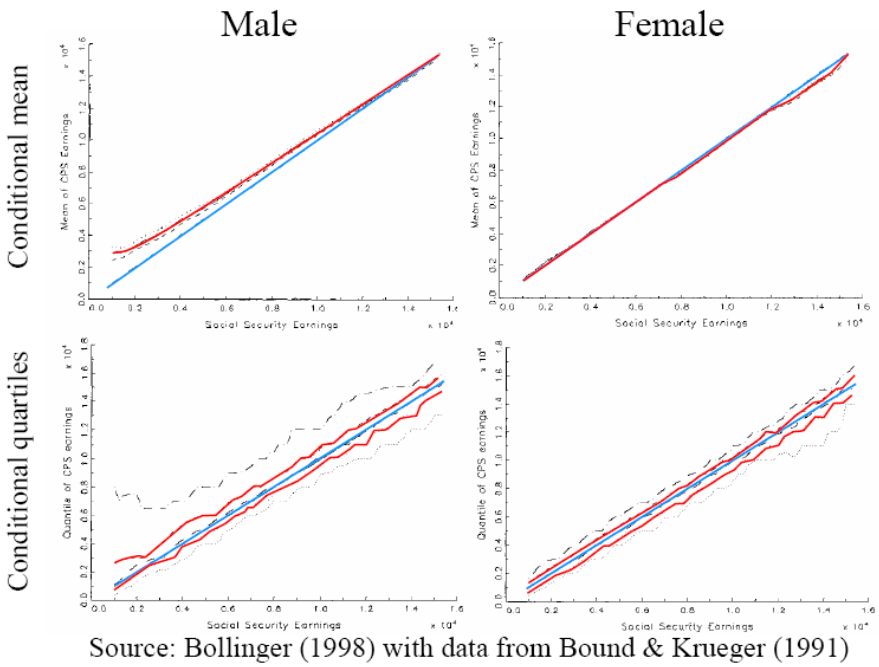


Figure 2.5: Self-reporting errors by gender

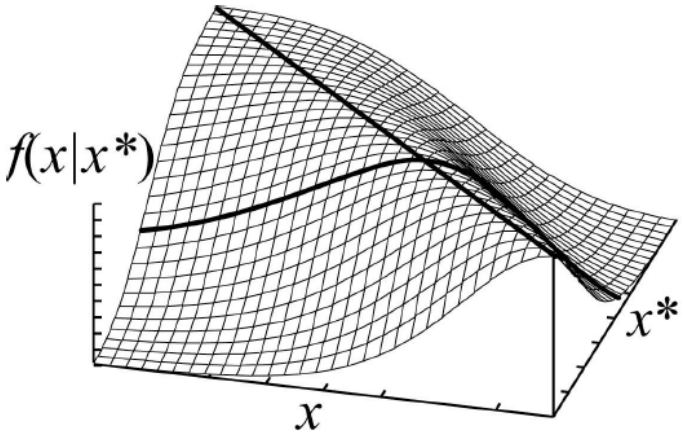


Figure 2.6: Graphical illustration of zero-mode measurement error

Theorem 2.4.2 *(Hu and Schennach (2008)) Under assumptions 2.4.5, 2.4.6, 2.4.7, and 2.4.8, the 2.1-measurement model in Definition 3 with a continuous X^* is non-parametrically identified in the sense that the joint distribution of the three variables (X, Y, Z), $f_{X,Y,Z}$, uniquely determines the joint distribution of the four variables (X, Y, Z, X^*), f_{X,Y,Z,X^*}, which satisfies*

$$f_{X,Y,Z,X^*} = f_{X|X^*} f_{Y|X^*} f_{Z|X^*} f_{X^*}. \tag{2.51}$$

This result implies that if we observe an additional binary indicator of the latent variable together with two measurements, we can relax the additivity and the independence assumptions in equation (2.30) and achieve nonparametric identification of very general models. Comparing the model in equation (2.30) and the 2.1-measurement model, which are both point identified, the latter is much more flexible to accommodate various economic models with latent variables. For example, Theorem 2.4.2 identifies the joint distribution of X^* and Z, and therefore, applies to both the case where $Z = X^* + \epsilon'$ and the case where the relationship between Z and X^* is specified as $X^* = Z + \epsilon'$. The latter case is related to the so-called Berkson-type measurement error models (Schennach (2013)).

Unlike the discrete case, it is difficult to exactly mimic the identification procedure to form an estimator in this general continuous case. However, the closed-form identification and estimation is possible under some nonparametric specification of the model, as shown in section 3.2.

2.4.5 An Illustrative Example

Here we use a simple example to illustrate the intuition of the identification results. Consider a labor supply model for college graduates, where Y is the 0-1 dichotomous employment status, X is the college GPA, Z is the SAT scores, and X^* is the latent ability type. We are interested in the probability of being employed given different ability, i.e., $\Pr(Y = 1|X^*)$, and the marginal probability of the latent ability f_{X^*}.

We consider a simplified version of the 2.1-measurement model with

$$\begin{aligned} \Pr(Y = 1|X^*) &\neq \Pr(Y = 1) \\ X &= X^*\gamma + \epsilon \\ Z &= X^*\gamma' + \epsilon' \end{aligned} \tag{2.52}$$

where $(X^*, \epsilon, \epsilon')$ are mutually independent. We may interpret the error term ϵ' as a performance shock in the SAT test. If coefficients γ and γ' are known, we can use X/γ and Z/γ' as the two measurements in equation (2.30) to identify the marginal distribution of ability without using the binary measurement Y. As shown in Hu and Sasaki (2015) , we can identify all the elements of interest in this model. Here we focus on the identification of the coefficients γ and γ' to illustrate the intuition of the identification results.

Since X^* is unobserved, we normalize $\gamma' = 1$ without loss of generality. A naive estimator for γ may be from the following regression equation

$$X = Z\gamma + (\epsilon - \epsilon'\gamma). \tag{2.53}$$

The OLS estimator corresponds to $\frac{cov(X,Z)}{var(Z)} = \gamma \frac{var(X^*)}{var(X^*)+var(\epsilon')}$, which is the well-known attenuation result with $|\frac{cov(X,Z)}{var(Z)}| < |\gamma|$. This regression equation suffers an endogeneity problem because the regressor, the SAT scores Z, does not perfectly reflect the ability X^* and is negatively correlated with the performance shock ϵ' in the regression error $(\epsilon - \epsilon'\gamma)$. When an additional variable Y is available, even if it is binary, we can use Y as an instrument to solve the endogeneity problem and identify γ as

$$\gamma = \frac{E[X|Y=1] - E[X|Y=0]}{E[Z|Y=1] - E[Z|Y=0]}. \tag{2.54}$$

This is literally the two-stage least square estimator. The regressor, SAT scores Z, is endogenous in both the employed subsample and the unemployed subsample. But the difference between the two subsamples may reveal how the observed GPA X is associated with ability X^* through γ.

The intuition of this identification strategy is that when we compare the employed $(Y = 1)$ subsample with the unemployed $(Y = 0)$ subsample, the only different element on the right hand side of the equation below is the marginal distribution of ability, i.e., $f_{X^*|Y=1}$ and $f_{X^*|Y=0}$ in

$$f_{X,Z|Y=y} = \int_{\mathcal{X}^*} f_{X|X^*} f_{Z|X^*} f_{X^*|Y=y} dx^*. \tag{2.55}$$

If we naively treat SAT scores Z as latent ability X^* to study the relationship between college GPA X and latent ability X^*, we may end up with a model with an endogeneity problem as in equation (2.53). However, the conditional independence assumption guarantees that the change in the employment status Y "exogenously" varies with latent ability X^*, and therefore, with the observed SAT scores Z, but does not vary with the performance shock ϵ', which is the cause of the endogeneity problem. Therefore, the employment status Y may serve as an instrument to achieve identification. Notice that this argument still holds if we compare the employed subsample with the whole sample, which is what we use in equations (2.48) and (2.49) in Section 2.4.3.[15]

Furthermore, an arguably surprising result is that such identification of the 2.1 measurement model is still nonparametric and global even if the instrument Y is binary. This is because the conditional independence assumption reduces the joint distribution f_{X,Y,Z,X^*} to distributions of each measurement conditional the latent variable $\left(f_{X|X^*}, f_{Y|X^*}, f_{Z|X^*}\right)$, and the marginal distribution f_{X^*} as in equation (2.42). The joint distribution f_{X,Y,Z,X^*} is a four-dimensional function, while $\left(f_{X|X^*}, f_{Y|X^*}, f_{Z|X^*}\right)$ are three two-dimensional functions. Therefore, the number of unknowns are greatly reduced under the conditional independence assumption.

[15] Another way to look at this is that γ can also be expressed as

$$\gamma = \frac{E[X|Y=1] - E[X]}{E[Z|Y=1] - E[Z]}.$$

2.5 A 3-measurement Model

We introduce the 2.1-measurement model to show the least data information needed for nonparametric identification of a measurement error model. Given that a random variable can always be transformed to a 0-1 dichotomous variable, the identification result can still hold when there are three measurements of the latent variable. In this section, we introduce the 3-measurement model to emphasize that three observables may play exchangeable roles so that it does not matter which measurement is called a dependent variable, a measurement, or an instrument variable. We define this case as follows:

Definition 4 *A **3-measurement model** contains three measurements, as in Definition 1, $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, and $Z \in \mathcal{Z}$ of the latent variable $X^* \in \mathcal{X}^*$ satisfying*

$$X \perp Y \perp Z \mid X^*, \tag{2.56}$$

i.e., (X, Y, Z) are jointly independent conditional on X^*.

Based on the results for the 2.1-measurement model, nonparametric identification of the joint distribution f_{X,Y,Z,X^*} in the 3-measurement model is feasible because one can always replace Y with a 0-1 binary indicator, e.g., $I(Y > E[Y])$. In fact, we intentionally write the results in section 2.4 in such a way that the assumptions and the theorems remain the same after replacing the binary support $\{0, 1\}$ with a general support $\mathcal{Y}$ for variable Y. An important observation here is that the three measurements (X, Y, Z) play exchangeable roles in the 3-measurement model. We can impose different restrictions on different measurements, which makes one look like a dependent variable, one like a measurement, and another like an instrument. But these "assignments" are arbitrary. On the one hand, the researcher may decide which "assignments" are reasonable based on the economic model. On the other hand, it does not matter which variable is called a dependent variable, a measurement, or an instrument variable in terms of identification. We summarize the results as follows:

Corollary 2.5.1 *Theorems 2.4.1 and 2.4.2 both hold for the 3-measurement model in Definition 4.*

2.6 A Measurement Model with 4 Observables

Before we extend the results to a dynamic setting, we introduce the nonparametric identification of measurement error models with two samples as in Carroll et al. (2010). Let Y be the dependent variable of interest, X^* be the latent explanatory variable, Z be other covariates, and $S \in \mathcal{S} = \{s_1, s_2\}$ be a sample indicator. The model of interest is described by $f_{Y|X^*,Z}$. The data structure can be described as follows:

Assumption 2.6.1 *i) conditional independence*

$$f_{Y,X|Z,S} = \sum_{X^*} f_{Y|X^*,Z} f_{X|X^*,S} f_{X^*|Z,S}. \tag{2.57}$$

ii) Y can be discretized to Y^d so that Y^d, X and X^* share the same support $\mathcal{X} = \{1, 2, ..., K\}$.

We define for any fixed (z, s)

$$\begin{aligned} M_{X,Y^d|z,s} &= \left[f_{X,Y^d|Z,S}(i,j|z,s)\right]_{i=1,2,...,K;j=1,2,...,K} \quad (2.58) \\ M_{X|X^*,s} &= \left[f_{X|X^*,S}(i|j,s)\right]_{i=1,2,...,K;j=1,2,...,K}. \\ D_{X^*|z,s} &= \begin{bmatrix} f_{X^*|Z,S}(1|z,s) & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & f_{X^*|Z,S}(K|z,s) \end{bmatrix} \quad (2.59) \end{aligned}$$

In general, we define a matrix representation of a probability distribution as follows: for discrete random variables R_1, R_2, R_3, the $(i+1, j+1)$-th element of the matrix M_{R_1,r_2,R_3} contains the joint probability that $(R_1 = i, R_2 = r_2, R_3 = j)$, for $i, j \in \{1, 2, ..., K\}$. Equation (2.57) then implies

$$M_{X,Y^d|z,s} = M_{X|X^*,s} D_{X^*|z,s} (M_{Y^d|X^*,z})^T. \quad (2.60)$$

We then assume that

Assumption 2.6.2 *Invertibility: for any $s \in \mathcal{S}$, there exists a $(z, \overline{z}, \overline{s})$ such that $M_{X,Y^d|z,s}$, $M_{X,Y^d|\overline{z},s}$, $M_{X,Y^d|\overline{z},\overline{s}}$, and $M_{X,Y^d|z,\overline{s}}$ are invertible and that for all $x^* \neq \widetilde{x}^*$ in $\mathcal{X}$*

$$\Delta_s \Delta_z \ln f_{X^*|Z,S}(x^*) \neq \Delta_s \Delta_z \ln f_{X^*|Z,S}(\widetilde{x}^*)$$

where $\Delta_s \Delta_z \ln f_{X^*|Z,S}(x^*)$ is defined as [16]

$$\begin{aligned} \Delta_s \Delta_z \ln f_{X^*|Z,S}(x^*) &= \left[\ln f_{X^*|Z,S}(x^*|z,s) - \ln f_{X^*|Z,S}(x^*|z,\overline{s})\right] \\ &\quad - \left[\ln f_{X^*|Z,S}(x^*|\overline{z},s) - \ln f_{X^*|Z,S}(x^*|\overline{z},\overline{s})\right]. \end{aligned}$$

Under the assumption that the four matrices on the LHS are invertible, which is directly testable, we may have

$$\begin{aligned} \mathbf{A} &\equiv M_{X,Y^d|z,s} M^{-1}_{X,Y^d|z,\overline{s}} \\ &= M_{X|X^*,s} D_{X^*|z,s} D^{-1}_{X^*|z,\overline{s}} M^{-1}_{X|X^*,\overline{s}}. \end{aligned}$$

Similar manipulations lead to

$$\begin{aligned} \mathbf{B} &\equiv M_{X,Y^d|\overline{z},\overline{s}} M^{-1}_{X,Y^d|\overline{z},s} \\ &= M_{X|X^*,\overline{s}} D_{X^*|\overline{z},\overline{s}} D^{-1}_{X^*|\overline{z},s} M^{-1}_{X|X^*,s}. \end{aligned}$$

[16] We use the ln function only for the purpose of using the double-difference notation.

Finally, we obtain

$$\begin{aligned} \mathbf{AB} &= M_{X|X^*,s} D_{X^*|z,s} D^{-1}_{X^*|z,\overline{s}} D_{X^*|\overline{z},\overline{s}} D^{-1}_{X^*|\overline{z},s} M^{-1}_{X|X^*,s}. \\ &\equiv M_{X|X^*,s} D_{z,\overline{z},s,\overline{s},X^*} M^{-1}_{X|X^*,s}, \end{aligned} \tag{2.61}$$

where

$$D_{z,\overline{z},s,\overline{s},X^*} = \begin{bmatrix} exp\left(\Delta_s \Delta_z \ln f_{X^*|Z,S}(1)\right) & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & exp\left(\Delta_s \Delta_z \ln f_{X^*|Z,S}(K)\right) \end{bmatrix} \tag{2.62}$$

Assumption 2.6.2 guarantees that this eigen-decomposition has distinctive eigenvalues. To pin down the ordering in one of the decompositions, we assume

Assumption 2.6.3 *There exists an $s_1 \in \mathcal{S}$ such that i) for any $s \in \mathcal{S}$, there exists a $(z, \overline{z}, s_1)$ satisfying Assumption 2.6.2; and ii) the ordering assumption 2.4.4 holds for $f_{X|X^*,S}(\cdot|\cdot, s_1)$.*

Therefore, we can identify $f_{X|X^*,S}(\cdot|\cdot, s_1)$. That identifies all the distributions $f_{X^*|Z,S}$ and $f_{Y|X^*,Z}$. We summarize the result as follows:

Theorem 2.6.1 *(Carroll, Chen and Hu, 2010) Under assumptions 2.6.1, 2.6.2, and 2.6.3, the conditional distribution of the two variables $f_{X,Y|Z,S}$ uniquely determines the conditional distribution of the three variables $f_{X,Y,X^*|Z,S}$ which satisfies*

$$f_{X,Y,X^*|Z,S} = f_{Y|X^*,Z} f_{X|X^*,S} f_{X^*|Z,S} \tag{2.63}$$

2.6.1 An Illustrative Example

In this section, we use a simple example to illustrate the identification strategy in Theorem 2.6.1, in Carroll et al. (2010) . Consider estimation of a consumption equation using two samples. Let Y be the consumption, X^* be the latent true income, Z be the family size, and $S \in \{s_1, s_2\}$ be a sample indicator. The data structure can be described as follows:

$$f_{Y,X|Z,S} = \int f_{Y|X^*,Z} f_{X|X^*,S} f_{X^*|Z,S} dx^*. \tag{2.64}$$

The consumption model is described by $f_{Y|X^*,Z}$, where consumption depends on income and family size. The self-reported income X may have different distributions in the two samples. The income X^* may be correlated with the family size Z and the income distribution may also be different in the two samples. Carroll et al. (2010) provide sufficient conditions for nonparametric identification of all the densities on the right hand side of equation (2.64).

To illustrate the identification strategy, we consider the following parametric specification

$$\begin{aligned} Y &= \beta X^* + \gamma Z + \eta \\ X &= X^* + \gamma' S + \epsilon \\ X^* &= \delta_1 S + \delta_2 Z + \delta_3 (S \times Z) + u, \end{aligned} \tag{2.65}$$

where $(\beta, \gamma, \gamma', \delta_1, \delta_2, \delta_3)$ are unknown constants with $\delta_3 \neq 0$.

We focus on the identification of β. If we naively treat X as the latent true income X^*, we have a model with endogeneity as follows:

$$\begin{aligned} Y &= \beta (X - \gamma' S - \epsilon) + \gamma Z + \eta \\ &= \beta X + \gamma Z - \beta\gamma' S + (\eta - \beta\epsilon). \end{aligned} \tag{2.66}$$

The regressor X is endogenous because it is correlated with the measurement error ϵ. Note that the income X^* may vary with the family size Z and the sample indicator S, which are independent of ϵ, the source of the endogeneity. The fact that there is no interaction term of Z and S on the right hand side of equation (2.66) is consistent with the conditional independence implied in equation (2.64). Let (z_0, z_1) and (s_0, s_1) be possible values of Z and S, respectively. Assuming $E[\eta|Z, S, X^*] = E[\epsilon|Z, S] = E[u|Z, S] = 0$, we estimate β as follows [17]

$$\beta = \frac{[E(Y|z_1, s_1) - E(Y|z_0, s_1)] - [E(Y|z_1, s_0) - E(Y|z_0, s_0)]}{[E(X|z_1, s_1) - E(X|z_0, s_1)] - [E(X|z_1, s_0) - E(X|z_0, s_0)]}. \tag{2.67}$$

This is a 2SLS estimator using $(S \times Z)$ as an IV in the first stage, in which the numerator is a difference-in-differences estimator for $\beta\delta_3 (z_1 - z_0)(s_1 - s_0)$ and the denominator is a difference-in-differences estimator for $\delta_3 (z_1 - z_0)(s_1 - s_0)$.

We may then consider two regressions

$$\begin{aligned} Y &= \beta(\delta_1 S + \delta_2 Z + \delta_3 (S \times Z) + u) + \gamma Z + \eta \\ &= \beta\delta_1 S + (\beta\delta_2 + \gamma) Z + \beta\delta_3 (S \times Z)) + (\eta + \beta u) \qquad (2.68) \\ X &= (\delta_1 S + \delta_2 Z + \delta_3 (S \times Z) + u) + \gamma' S + \epsilon \\ &= (\delta_1 + \gamma') S + \delta_2 Z + \delta_3 (S \times Z) + (\epsilon + u), \qquad (2.69) \end{aligned}$$

Therefore, we can estimate all the coefficients from these two regressions. Furthermore, we can extend this results to the case where there are other covariates W as follows:

$$\begin{aligned} Y &= \beta X^* + \gamma Z + \theta_y W + \eta \\ X &= X^* + \gamma' S + \theta_x W + \epsilon \\ X^* &= \delta_1 S + \delta_2 Z + \delta_3 (S \times Z) + \theta W + u, \end{aligned} \tag{2.70}$$

[17] Due to the linear specification of the model, the result here remains with $E[\eta|Z, S] = 0$ instead of $E[\eta|Z, S, X^*] = 0$. That means X^* is allowed to be endogenous in the regression equation. In other words, the estimator can tackle both endogeneity and measurement error issues in some settings.

The regression equations with observables are

$$\begin{aligned} Y &= \beta(\delta_1 S + \delta_2 Z + \delta_3 (S \times Z) + \theta W + u) + \gamma Z + \theta_y W + \eta \\ &= \beta\delta_1 S + (\beta\delta_2 + \gamma) Z + \beta\delta_3 (S \times Z)) + (\beta\theta + \theta_y) W + (\eta + \beta u) \qquad (2.71) \\ X &= (\delta_1 S + \delta_2 Z + \delta_3 (S \times Z) + \theta W + u) + \gamma' S + \theta_x W + \epsilon \\ &= (\delta_1 + \gamma') S + \delta_2 Z + \delta_3 (S \times Z) + (\theta + \theta_x) W + (\epsilon + u). \qquad (2.72) \end{aligned}$$

We can identify $(\beta\theta + \theta_y)$ and $(\theta + \theta_x)$, besides other parameters. Therefore, we can identify the coefficients on those covariates which don't appear in one of the three equations, or appear with a normalized coefficients.

2.7 Dynamic Measurement Models

2.7.1 Hidden Markov Models

The 3-measurement model is directly applicable to identify a hidden Markov model. We consider a hidden Markov model containing $\{X_t, X_t^*\}$, where $\{X_t^*\}$ is a latent first-order Markov process, i.e.,

$$X_{t+1}^* \perp \{X_s^*\}_{s \leq t-1} \mid X_t^*. \qquad (2.73)$$

In each period, we observe a measurement X_t of the latent X_t^* satisfying

$$X_t \perp \{X_s, X_s^*\}_{s \neq t} \mid X_t^*. \qquad (2.74)$$

This is the so-called local independence assumption, where a measurement X_t is independent of everything else conditional the latent variable X_t^* in the sample period. The relationship among the variables can be shown in the flow chart as follows.

$$\begin{array}{ccccccc} & & X_{t-1} & & X_t & & X_{t+1} \\ & & \uparrow & & \uparrow & & \uparrow \\ \longrightarrow & X_{t-1}^* & \longrightarrow & X_t^* & \longrightarrow & X_{t+1}^* & \longrightarrow \end{array}$$

Consider a panel data set, where we observed three periods of data $\{X_{t-1}, X_t, X_{t+1}\}$. The conditions in equations (2.73) and (2.74) imply

$$X_{t-1} \perp X_t \perp X_{t+1} \mid X_t^*, \qquad (2.75)$$

i.e., (X_{t-1}, X_t, X_{t+1}) are jointly independent conditional on X_t^*. Although the original model is dynamic, it can be reduced to a 3-measurement model as in equation (2.75). Corollary 2.5.1 then non-parametrically identifies $f_{X_{t+1}|X_t^*}$, $f_{X_t|X_t^*}$, $f_{X_{t-1}|X_t^*}$, and $f_{X_t^*}$. Under a stationarity assumption that $f_{X_{t+1}|X_{t+1}^*} = f_{X_t|X_t^*}$, we can then identify the Markov kernel $f_{X_{t+1}^*|X_t^*}$ from

$$f_{X_{t+1}|X_t^*} = \int_{\mathcal{X}^*} f_{X_{t+1}|X_{t+1}^*} f_{X_{t+1}^*|X_t^*} dx_{t+1}^*, \qquad (2.76)$$

by inverting the integral operator corresponding to $f_{X_{t+1}|X^*_{t+1}}$.[18] Therefore, it does not really matter which one of $\{X_{t-1}, X_t, X_{t+1}\}$ is treated as measurement or instrument for X^*_t. Applications of nonparametric identification of such a hidden Markov model or, in general, the 3-measurement model based on Hu (2008) and Hu and Schennach (2008), can be found in Hu et al. (2013b) , Feng and Hu (2013) , Wilhelm (2013) , and Hu and Sasaki (2018) , etc.

2.7.2 Markov Models with Limited Feedback

A natural extension to the hidden Markov model in equations (2.73)-(2.74) is to relax the local independence assumption in equation (2.74) when more periods of data are available. For example, we may allow direct serial correlation among observed measurements $\{X_t\}$ of latent variables $\{X^*_t\}$. To this end, we assume the following:

Assumption 2.7.1 *The joint process $\{X_t, X^*_t\}$ is a first-order Markov process. Furthermore, the Markov kernel satisfies*

$$f_{X_t,X^*_t|X_{t-1},X^*_{t-1}} = f_{X_t|X^*_t,X_{t-1}} f_{X^*_t|X_{t-1},X^*_{t-1}}. \tag{2.77}$$

Equation (2.77) is the so-called limited feedback assumption in Hu and Shum (2012) . It implies that the latent variable in current period has summarized all the information on the latent part of the process. The relationship among the variables may be described as follows:

$$\begin{array}{ccccccccccc} \longrightarrow & & X_{t-2} & \longrightarrow & X_{t-1} & \longrightarrow & X_t & \longrightarrow & X_{t+1} & \longrightarrow \\ \searrow & & \updownarrow & \searrow & \updownarrow & \searrow & \updownarrow & \searrow & \updownarrow & \searrow \\ \longrightarrow & & X^*_{t-2} & \longrightarrow & X^*_{t-1} & \longrightarrow & X^*_t & \longrightarrow & X^*_{t+1} & \longrightarrow \end{array}$$

For simplicity, we focus on the discrete case and assume

Assumption 2.7.2 *X_t and X^*_t share the same support $\mathcal{X}^* = \{x^*_1, x^*_2, \ldots, x^*_K\}$.*

The observed distribution is associated with unobserved ones as follows:

$$f_{X_{t+1},X_t,X_{t-1},X_{t-2}} = \sum_{x^*} f_{X_{t+1}|X_t,X^*_t} f_{X_t|X^*_t,X_{t-1}} f_{X^*_t,X_{t-1},X_{t-2}}. \tag{2.78}$$

We define for any fixed (x_t, x_{t-1})

$$\begin{aligned} M_{X_{t+1},x_t|x_{t-1},X_{t-2}} &= \left[f_{X_{t+1},X_t|X_{t-1},X_{t-2}}(x_i, x_t|x_{t-1}, x_j)\right]_{i=1,2,\ldots,K;j=1,2,\ldots,K} \\ M_{X_t|x_{t-1},X_{t-2}} &= \left[f_{X_t|X_{t-1},X_{t-2}}(x_i|x_{t-1}, x_j)\right]_{i=1,2,\ldots,K;j=1,2,\ldots,K}. \end{aligned} \tag{2.79}$$

Assumption 2.7.3 *(i) for any $x_{t-1} \in \mathcal{X}$, $M_{X_t|x_{t-1},X_{t-2}}$ is invertible.*

[18] Without stationarity, one can use one more period of data, i.e., X_{t+2}, to identify $f_{X_{t+1}|X^*_{t+1}}$ from the joint distribution of (X_t, X_{t+1}, X_{t+2}) .

(ii) for any $x_t \in \mathcal{X}$*, there exists a* $(x_{t-1}, \overline{x}_{t-1}, \overline{x}_t)$ *such that* $M_{X_{t+1},x_t|x_{t-1},X_{t-2}}$, $M_{X_{t+1},x_t|\overline{x}_{t-1},X_{t-2}}$, $M_{X_{t+1},\overline{x}_t|x_{t-1},X_{t-2}}$, *and* $M_{X_{t+1},\overline{x}_t|\overline{x}_{t-1},X_{t-2}}$ *are invertible and that for all* $x_t^* \neq \widetilde{x}_t^*$ *in* $\mathcal{X}^*$

$$\Delta_{x_t}\Delta_{x_{t-1}} \ln f_{X_t|X_t^*,X_{t-1}}(x_t^*) \neq \Delta_{x_t}\Delta_{x_{t-1}} \ln f_{X_t|X_t^*,X_{t-1}}(\widetilde{x}_t^*)$$

where $\Delta_{x_t}\Delta_{x_{t-1}} \ln f_{X_t|X_t^*,X_{t-1}}(x_t^*)$ *is defined as*

$$\begin{aligned}\Delta_{x_t}\Delta_{x_{t-1}} \ln f_{X_t|X_t^*,X_{t-1}}(x_t^*) \;:=\; &\left[\ln f_{X_t|X_t^*,X_{t-1}}(x_t|x_t^*,x_{t-1}) - \ln f_{X_t|X_t^*,X_{t-1}}(x_t|x_t^*,\overline{x}_{t-1})\right] \\ &- \left[\ln f_{X_t|X_t^*,X_{t-1}}(\overline{x}_t|x_t^*,x_{t-1}) - \ln f_{X_t|X_t^*,X_{t-1}}(\overline{x}_t|x_t^*,\overline{x}_{t-1})\right].\end{aligned}$$

Assumption 2.7.4 *For any* $x_t \in \mathcal{X}$*, there exists a known functional* M *such that* $M\left[f_{X_{t+1}|X_t,X_t^*}(\cdot|x_t,x_t^*)\right]$ *is strictly increasing in* x_t^**.*

Assumption 2.7.5 *The Markov kernel is stationary, i.e.,*

$$f_{X_t,X_t^*|X_{t-1},X_{t-1}^*} = f_{X_2,X_2^*|X_1,X_1^*}. \tag{2.80}$$

The invertibility in Assumption 2.7.3 is testable because it imposes a rank condition on observed matrices. The invertibility guarantees that a directly estimable matrix has an eigenvalue-eigenvector decomposition, where the eigenvalues are associated with $\Delta_{x_t}\Delta_{x_{t-1}} \ln f_{X_t|X_t^*,X_{t-1}}$ and the eigenvectors are related to $f_{X_{t+1}|X_t,X_t^*}(\cdot|x_t,x_t^*)$ for a fixed x_t. Assumption 2.7.3(ii) is needed for the distinctiveness of the eigenvalues. And Assumption 2.7.4 reveals the ordering of the eigenvectors as Assumption 2.4.8. Assumption 2.7.5 is a stationarity assumption, which is not needed with one more periods of data. We summarize the results as follows:

Theorem 2.7.1 *(Hu and Shum (2012)) Under assumptions 2.7.1, 2.7.2, 2.7.3, 2.7.4, and 2.7.5, the joint distribution of four periods of data* $f_{X_{t+1},X_t,X_{t-1},X_{t-2}}$ *uniquely determines the Markov transition kernel* $f_{X_t,X_t^*|X_{t-1},X_{t-1}^*}$ *and the initial condition* f_{X_{t-2},X_{t-2}^*}*.*

Proof: See section 6.1 for details.

For the continuous case and other variations of the assumptions, such as non-stationarity, I refer to Hu and Shum (2012) for details. A simple extension of this result is the case where X_t^* is discrete and X_t is continuous. As in the discussion following Theorem 2.4.1, the identification results still apply with minor modification of the assumptions.

In the case where $X_t^* = X^*$ is time-invariant, the condition in equation (2.77) is not restrictive and the Markov kernel becomes $f_{X_t|X_{t-1},X^*}$. For such a first-order Markov model, Kasahara and Shimotsu (2009) suggest using two periods of data to break the interdependence and use six periods of data to identify the transition kernel. For fixed $X_t = x_t$, $X_{t+2} = x_{t+2}$, $X_{t+4} = x_{t+4}$, it can be shown that $X_{t+1}, X_{t+3}, X_{t+5}$ are independent conditional on X^* as follows:

$$f_{X_{t+5},x_{t+4},X_{t+3},x_{t+2},X_{t+1},x_t} = \sum_{x^* \in \mathcal{X}^*} f_{X_{t+5}|x_{t+4},X^*} f_{x_{t+4},X_{t+3}|x_{t+2},X^*} f_{x_{t+2}.X_{t+1},x_t,X^*}.$$

The model then falls into the framework of the 3-measurement model, where $(X_{t+1}, X_{t+3}, X_{t+5})$ may serve as three measurements for each fixed (x_t, x_{t+2}, x_{t+4}) to achieve identification. [19] This similarity to the 3-measurement model can also be seen in Bonhomme et al. (2016a) and Bonhomme et al. (2016b) . However, the 2.1-measurement model implies that minimum data information for nonparametric identification can be "2.1 measurements" instead of "3 measurements". Hu and Shum (2012) shows that the interaction between observables in the middle two periods may play the role of the binary measurement in the 2.1-measurement model so that such a model, even with a time-varying unobserved state variable, can be identified using only four periods of data.

See section 6.1 for detailed description of this model with applications to dynamic discrete choice models.

2.7.3 An Illustrative Example

In the dynamic model in Theorem 2.7.1, we can re-write equation (2.78) as

$$f_{X_{t+1},X_{t-2}|X_t,X_{t-1}} = \sum_{x^*} f_{X_{t+1}|X_t^*,X_t} f_{X_{t-2}|X_t^*,X_{t-1}} f_{X_t^*|X_t,X_{t-1}}, \tag{2.81}$$

which is analogical to equation (2.64). Similar to the previous example on consumption, suppose we naively treat X_{t-2} as X_t^* to study the relationship between X_{t+1} and (X_t, X_t^*), say $X_{t+1} = H(X_t^*, X_t, \eta_t)$, where η_t is an independent error term. And suppose the conditional density $f_{X_{t-2}|X_t^*,X_{t-1}}$ implies $X_{t-2} = G(X_t^*, X_{t-1}, \epsilon_t)$, where ϵ_t represents an independent error term. Suppose we can replace X_t^* by $G^{-1}(X_{t-2}, X_{t-1}, \epsilon_t)$ to obtain

$$X_{t+1} = H\left(G^{-1}(X_{t-2}, X_{t-1}, \epsilon_t), X_t, \eta_t\right), \tag{2.82}$$

where X_{t-2} is endogenous and correlated with ϵ. The conditional independence in equation (2.81) implies that the variation in X_t and X_{t-1} may vary with X_t^*, but not with the error ϵ. However, the variation in X_t may change the relationship between the future X_{t+1} and the latent variable X_t^*, while the variation in X_{t-1} may change the relationship between the early X_{t-2} and the latent X_t^*. Therefore, a "joint" second-order variation in (X_t, X_{t-1}) may lead to an "exogenous" variation in X^*, which may solve the endogeneity problem. Thus, our identification strategy may be considered as a nonparametric version of a difference-in-differences argument.

For example, let X_t stand for the choice of health insurance between a high coverage plan and a low coverage plan. And X_t^* stands for the good or bad health status. The Markov process $\{X_t, X_t^*\}$ describes the interaction between insurance choices and health status. We consider the joint distribution of four periods of insurance choices $f_{X_{t+1},X_t,X_{t-1},X_{t-2}}$. If we compare a subsample with (X_t, X_{t-1}) = (high, high) and a subsample with and (X_t, X_{t-1}) = (high, low), we should be able to "difference out" the direct impact of health insurance choice X_t on the choice X_{t+1} in next period in $f_{X_{t+1}|X_t^*,X_t}$. Then, we may repeat such a comparison again with (X_t, X_{t-1}) = (low, high) and (X_t, X_{t-1}) = (low, low). In both

[19] See section 6.1.3 for detailed comparison with Kasahara and Shimotsu (2009).

comparisons, the impact of changes in insurance choice X_{t-1} described in $f_{X_{t-2}|X_t^*,X_{t-1}}$ is independent of the choice X_t. Therefore, the difference in the differences from those two comparisons above may lead to exogenous variation in X_t^* as described in $f_{X_t^*|X_t,X_{t-1}}$, which is independent of the endogenous error due to naively using X_{t-2} as X_t^*. Therefore, the second-order joint variation in observed insurance choices (X_t, X_{t-1}) may serve as an instrument to solve the endogeneity problem caused by using the observed insurance choice X_{t-2} as a proxy for the unobserved health condition X_t^*.

3

Semiparametric and Nonparametric Estimation

This paper focuses on nonparametric identification of models with latent variables and its applications in applied microeconomic models. Given the length limit of the paper, I only provide a brief description of estimators proposed for the models above. All the identification results above are at the distribution level in the sense that probability distribution functions involving latent variables are uniquely determined by probability distribution functions of observables, which are directly estimable from a random sample of observables. Therefore, a maximum likelihood estimator is a straightforward choice for these models.

3.1 Sieve Maximum Likelihood Estimators

Consider the 2.1-measurement model in Theorem 2.4.2, where the observed density is associated with the unobserved ones as follows:

$$f_{X,Y,Z}(x,y,z) = \int_{\mathcal{X}^*} f_{X|X^*}(x|x^*) f_{Y|X^*}(y|x^*) f_{Z|X^*}(z|x^*) f_{X^*}(x^*) dx^*. \tag{3.1}$$

Our identification results provide conditions under which this equation has a unique solution $\left(f_{X|X^*}, f_{Y|X^*}, f_{Z|X^*}, f_{X^*}\right)$. Suppose that Y is the dependent variable and the model of interest is described by a parametric conditional density function as

$$f_{Y|X^*}(y|x^*) = f_{Y|X^*}(y|x^*;\theta). \tag{3.2}$$

With an i.i.d. sample $\{x_i, y_i, z_i\}_{i=1,2,...,N}$, we can use a sieve maximum likelihood estimator (Shen (1997) and Chen and Shen (1998)) based on

$$\left(\widehat{\theta}, \widehat{f}_{x|x^*}, \widehat{f}_{z|x^*}, \widehat{f}_{x^*}\right) = \underset{(\theta,f_1,f_2,f_3)\in\mathcal{A}_N}{\arg\max} \frac{1}{N}\sum_{i=1}^{N} \ln \int_{\mathcal{X}^*} f_1(x_i|x^*) f_{Y|X^*}(y_i|x^*;\theta) f_2(z_i|x^*) f_3(x^*) dx^*, \tag{3.3}$$

where $\mathcal{A}_N$ is approximating sieve spaces which contain truncated series as parametric approximations to densities $\left(f_{X|X^*}, f_{Z|X^*}, f_{X^*}\right)$. For example, function $f_1(x|x^*)$ in the sieve

space $\mathcal{A}_N$ can be as follows:

$$f_1(x|x^*) = \sum_{j=1}^{J_N} \sum_{k=1}^{K_N} \beta_{jk} p_j\left(x - x^*\right) p_k\left(x^*\right),$$

where $p_j(\cdot)$ is a known basis function, such as power series, splines, Fourier series, etc. and J_N and K_N are smoothing parameters. The choice of a sieve space depends on how well it can approximate the original functional space and how much computation burden it may lead to (See section 2.3.6 of Chen (2007) for details). One advantage of a sieve estimator is that it is relatively convenient to impose restrictions on the sieve space $\mathcal{A}_N$. To be specific, Assumption 2.4.8 can be imposed on the sieve coefficients β_{jk} (See section S4 of supplementary materials of Hu and Schennach (2008) for details). Since the coefficients are treated as unknown parameters in the likelihood function, the parameters of interest in Equation (3.3) can be estimated just as a parametric MLE. The number of coefficients $J_N \times K_N$ diverges at a given speed with the sample size N, which makes the approximation more flexible with a larger sample size. A useful result worth mentioning is that the parametric part of the model can converge at a fast rate, i.e., $\widehat{\theta}$ can be $\sqrt{n}$ consistent and asymptotically normally distributed under suitable assumptions (Shen (1997)). We refer to Hu and Schennach (2008) , Carroll et al. (2010) and supplementary materials for more discussion on this semi-nonparametric extremum estimator.

3.1.1 A Setup

Given the general nonparametric identification, we develop our estimator based on an i.i.d sample, which can be extended to for time series data. We assume that there is a random sample $\{x_i, y_i, z_i\}_{i=1}^n$.

We adopt a parametric specification in equation (3.2) and leave other elements nonparametrically. Let the true value of the unknowns be $\alpha_0 \equiv (\theta_0^T, f_{01}, f_{02}, f_{03})^T \equiv (\theta_0^T, f_{x|x^*}, f_{z|x^*}, f_{x^*})^T$, where $f_{A|B}$ denotes the distribution of A conditional on B. We then introduce a sieve MLE estimator $\widehat{\alpha}$ for α_0, and establish the asymptotic normality of $\widehat{\theta}$. These results can also be extended to the case where the function m is misspecified.

Following Hu and Schennach (2008) and Carroll et al. (2010), we consider the widely used Hölder space of functions. Let $\xi = (\xi_1, \xi_2, \xi_3)^T \in \mathbb{R}^3$, $a = (a_1, a_2, a_3)^T$, and $\nabla^a h(\xi) \equiv \frac{\partial^{a_1+a_2+a_3} h(\xi_1,\xi_2,\xi_3)}{\partial \xi_1^{a_1} \partial \xi_2^{a_2} \partial \xi_3^{a_3}}$ denote the $(a_1 + a_2 + a_3)^{\text{th}}$ derivative. Let $\|\cdot\|_E$ denote the Euclidean norm. Let $\mathcal{V} \subseteq \mathbb{R}^3$ and $\underline{\gamma}$ be the largest integer satisfying $\gamma > \underline{\gamma}$. The Hölder space $\Lambda^\gamma(\mathcal{V})$ of order $\gamma > 0$ is a space of functions $h : \mathcal{V} \mapsto \mathbb{R}$, such that the first $\underline{\gamma}$ derivatives are continuous and bounded, and the $\underline{\gamma}^{\text{th}}$ derivative is Hölder continuous with the exponent $\gamma - \underline{\gamma} \in (0, 1]$. We define a Hölder ball as $\Lambda_c^\gamma(\mathcal{V}) \equiv \{h \in \Lambda^\gamma(\mathcal{V}) : \|h\|_{\Lambda^\gamma} \leq c < \infty\}$, in which

$$\|h\|_{\Lambda^\gamma} \equiv \max_{a_1+a_2+a_3 \leq \underline{\gamma}} \sup_{\xi} \left|\nabla^a h(\xi)\right| + \max_{a_1+a_2+a_3 = \underline{\gamma}} \sup_{\xi \neq \xi'} \frac{\left|\nabla^a h(\xi) - \nabla^a h(\xi')\right|}{\left(\|\xi - \xi'\|_E\right)^{\gamma - \underline{\gamma}}} < \infty.$$

The space containing $f_{01} = f_{x|x^*}$ are assumed to be

$$\mathcal{F}_1 = \left\{ \begin{array}{c} f_1(\cdot|\cdot) \in \Lambda_c^{\gamma_1}(\mathcal{X} \times \mathcal{X}^*) : \text{Assumptions in Theorem 2.4.2 hold,} \\ f_1(\cdot|x^*) \text{ is a positive density function for all } x^* \in \mathcal{X}^* \end{array} \right\}.$$

Similarly, we assume f_{02} and f_{03} are in the following functional spaces

$$\mathcal{F}_2 = \left\{ \begin{array}{c} f_2(\cdot|\cdot) \in \Lambda_c^{\gamma_2}(\mathcal{Z} \times \mathcal{X}^*) : \text{Assumptions in Theorem 2.4.2 hold,} \\ f_2(\cdot|x^*) \text{ is a positive density function for all } x^* \in \mathcal{X}^* \end{array} \right\},$$

and

$$\mathcal{F}_3 = \{ f_3(\cdot) \in \Lambda_c^{\gamma_3}(\mathcal{X}^*) : f_3(\cdot) \text{ is a positive density function } \}.$$

Let $\mathcal{A} = \Theta \times \mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3$ as the parameter space. The log-joint likelihood for $\alpha \equiv (\theta^T, f_1, f_2, f_3)^T \in \mathcal{A}$ is given by:

$$\sum_{i=1}^{n} \log f(x_i, y_i, z_i) = \sum_{i=1}^{n} \ell(D_i; \alpha),$$

in which $D_i = (x_i, y_i, z_i)$ and

$$\begin{aligned} \ell(D_i; \alpha) &\equiv \ell(x_i, y_i, z_i; \theta, f_1, f_2, f_3) \\ &= \log\{ \int f_1(x_i|x^*) f_{y|x^*}(y_i|x^*; \theta) f_2(z_i|x^*) f_3(x^*) dx^* \}. \end{aligned}$$

Let $E[\cdot]$ denote the expectation with respect to the underlying true data generating process for D_i. Then

$$\alpha_0 = \arg\sup_{\alpha \in \mathcal{A}} E\left[\ell(D_i; \alpha)\right].$$

We then use a sequence of finite-dimensional sieve spaces $\mathcal{A}_n = \Theta \times \mathcal{F}_1^n \times \mathcal{F}_2^n \times \mathcal{F}_3^n$ to approximate the functional space $\mathcal{A} = \Theta \times \mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3$. The semiparametric sieve MLE $\widehat{\alpha}_n = \left(\widehat{\theta}^T, \widehat{f}_1, \widehat{f}_2, \widehat{f}_3\right)^T \in \mathcal{A}_n$ for $\alpha_0 \in \mathcal{A}$ is defined as:

$$\widehat{\alpha}_n = \arg\max_{\alpha \in \mathcal{A}_n} \sum_{i=1}^{n} \ell(D_i; \alpha).$$

Let $p^{k_n}(\cdot)$ be a $k_n \times 1-$vector of known basis functions, such as power series, splines, Fourier series, Legendre polynomials, Hermite polynomials, etc. We use linear sieves to directly approximate unknown densities:

$$\mathcal{F}_1^n = \left\{ f_1(x|x^*) = p^{k_{1,n}}(x - x^*)^T \left[\beta_{1,i,j}\right]_{i,j} p^{k_{1,n}}(x^*) \in \mathcal{F}_1 \right\}$$

$$\mathcal{F}_2^n = \left\{ f_2(z|x^*) = \left[p^{k_{2,n}}(z)^T \left[\beta_{2,i,j}\right]_{i,j} p^{k_{2,n}}(x^*) \right]^2 \in \mathcal{F}_2 \right\}$$

$$\mathcal{F}_3^n = \left\{ f_3(x^*) = \left[p^{k_{3,n}}(x^*)^T \beta_3 \right]^2 \in \mathcal{F}_3 \right\}$$

where $\left[\beta_{1,i,j}\right]_{i,j}$ and $\left[\beta_{2,i,j}\right]_{i,j}$ represent matrices of sieve coefficients. Below we present the asymptotic properties of the proposed estimator.

3.1.2 Consistency

Here we provide sufficient conditions for the consistency of the sieve estimator $\widehat{\alpha}_n = \left(\widehat{\theta}^T, \widehat{f}_1, \widehat{f}_2, \widehat{f}_3, \right)^T$.

Assumption 3.1.1 *(i) All the assumptions in Theorem 2.4.2 hold; (ii)* $f_{x|x^*}(\cdot|\cdot) \in \mathcal{F}_1$ *with* $\gamma_1 > 1/2$ *; (iii)* $f_{z|x^*}(\cdot|\cdot) \in \mathcal{F}_2$ *with* $\gamma_2 > 1$*; (iv)* $f_{x^*}(\cdot) \in \mathcal{F}_3$ *with* $\gamma_3 > 1$.

Assumption 3.1.2 *(i)* $\{x_i, y_i, z_i\}_{i=1}^n$ *is i.i.d.; (ii)* $f(y|x^*;\theta)$ *is continuous in* $\theta \in \Theta$*, and* Θ *is a compact subset of* $\mathbb{R}^{d_\theta}$*; (iii)* $\theta_0 \in \Theta$ *is the unique solution of* $max_\theta E[lnf(y|x^*;\theta)]$ *over* $\theta \in \Theta$.

We define a norm on $\mathcal{A}$ as: $\|\alpha\|_s = \|\theta\|_E + \|f_1\|_{\infty,\omega_1} + \|f_2\|_{\infty,\omega_2} + \|f_3\|_{\infty,\omega_3}$ in which $\|h\|_{\infty,\omega_j} \equiv \sup_\xi |h(\xi)\omega_j(\xi)|$ with $\omega_j(\xi) = \left(1 + \|\xi\|_E^2\right)^{-\varsigma_j/2}$, $\varsigma_j > 0$ for $j = 1, 2, 3$. We assume

Assumption 3.1.3 *(i)* $-\infty < E[\ell(D_i;\alpha_0)] < \infty$*,* $E[\ell(D_i;\alpha)]$ *is upper semicontinuous on* $\mathcal{A}$ *under the metric* $\|\cdot\|_s$*; (ii) there is a finite* $\tau > 0$ *and a random variable* $U(D_i)$ *with* $E\{U(D_i)\} < \infty$ *such that* $\sup_{\alpha \in \mathcal{A}_n : \|\alpha - \alpha_0\|_s \leq \delta} |\ell(D_i;\alpha) - \ell(D_i;\alpha_0)| \leq \delta^\tau U(D_i)$.

Assumption 3.1.4 *(i)* $p^{k_{j,n}}(\cdot)$ *is a* $k_{j,n} \times 1-$*vector of basis functions on* $\mathbb{R}$ *for* $j = 1, 2, 3$*; (ii)* $\min\{k_{1,n}^2, k_{2,n}^2, k_{3,n}\} \to \infty$ *and* $\max\{k_{1,n}^2, k_{2,n}^2, k_{3,n}\}/n \to 0$.

We then have

Lemma 3.1.1 Under Assumptions 3.1.1–3.1.4, we have $\|\widehat{\alpha}_n - \alpha_0\|_s = o_p(1)$.

This is a direct extension from Carroll et al. (2010) , which uses theorem 3.1 in Chen (2007).

3.1.3 Convergence Rates and Asymptotic Normality

The asymptotic properties of our estimator is a direct extension of that in Carroll et al. (2010). We list the conditions below for readers' convenience.

Convergence Rates of Nonparametric Part

Given the consistency shown in Lemma 3.1.1, we focus on a shrinking $||\cdot||_s-$neighborhood around α_0. Let $\mathcal{A}_{0s} \equiv \{\alpha \in \mathcal{A} : ||\alpha - \alpha_0||_s = o(1), ||\alpha||_s \leq c_0 < c\}$ and $\mathcal{A}_{0sn} \equiv \{\alpha \in \mathcal{A}_n : ||\alpha - \alpha_0||_s = o(1), ||\alpha||_s \leq c_0 < c\}$. We assume that both $\mathcal{A}_{0s}$ and $\mathcal{A}_{0sn}$ are convex parameter spaces, and that $\ell(D_i;\alpha + \tau v)$ is twice continuously differentiable at $\tau = 0$ for almost all D_i and any direction $v \in \mathcal{A}_{0s}$.

Define the pathwise first and second derivatives of the sieve loglikelihood in the direction v as

$$\frac{d\ell(D_i;\alpha)}{d\alpha}[v] \equiv \frac{d\ell(D_i;\alpha+\tau v)}{d\tau}|_{\tau=0}; \qquad \frac{d^2\ell(D_i;\alpha)}{d\alpha d\alpha^T}[v,v] \equiv \frac{d^2\ell(D_i;\alpha+\tau v)}{d\tau^2}|_{\tau=0}.$$

Mimicing Ai and Chen (2007), for any $\alpha_1, \alpha_2 \in \mathcal{A}_{0s}$, we define a pseudo metric $||\cdot||_2$ as

$$\|\alpha_1 - \alpha_2\|_2 \equiv \sqrt{-E\left(\frac{d^2\ell(D_i;\alpha_0)}{d\alpha d\alpha^T}[\alpha_1-\alpha_2, \alpha_1-\alpha_2]\right)}.$$

Our goal is to show that $\widehat{\alpha}_n$ converges to α_0 at a rate faster than $n^{-1/4}$ under the pseudo metric $\|\cdot\|_2$. We make the following assumptions:

Assumption 3.1.5 *(i) $\varsigma_j > \gamma_j$ for $j = 1, 2, 3$; (ii) $\max\{k_{1,n}^{-\gamma_1}, k_{2,n}^{-\gamma_2}, k_{3,n}^{-\gamma_3}\} = o(n^{-1/4})$.*

Assumption 3.1.6 *(i) $\mathcal{A}_{0s}$ is convex at α_0 and $\theta_0 \in int\,(\Theta)$; (ii) $\ell(D_i;\alpha)$ is twice continuously pathwise differentiable with respect to $\alpha \in \mathcal{A}_{0s}$, and $m(y^*;\theta)$ is twice continuously differentiable at θ_0.*

Assumption 3.1.7 $\sup_{\widetilde{\alpha}\in\mathcal{A}_{0s}} \sup_{\alpha\in\mathcal{A}_{0sn}} \left|\frac{d\ell(D_i;\widetilde{\alpha})}{d\alpha}\left[\frac{\alpha-\alpha_0}{\|\alpha-\alpha_0\|_s}\right]\right| \leq U(D_i)$ *for a random variable $U(D_i)$ with $E\{[U(D_i)]^2\} < \infty$.*

Assumption 3.1.8 *(i)* $\sup_{v\in\mathcal{A}_{0s}:||v||_s=1} -E\left(\frac{d^2\ell(D_i;\alpha_0)}{d\alpha d\alpha^T}[v,v]\right) \leq C < \infty$*; (ii) uniformly over $\widetilde{\alpha} \in \mathcal{A}_{0s}$ and $\alpha \in \mathcal{A}_{0sn}$, we have*

$$-E\left(\frac{d^2\ell(D_i;\widetilde{\alpha})}{d\alpha d\alpha^T}[\alpha-\alpha_0, \alpha-\alpha_0]\right) = \|\alpha-\alpha_0\|_2^2 \times \{1+o(1)\}.$$

These assumptions are standard in the literature. As a direct application of Theorem 3.2 of Shen and Wong (1994) to the local parameter space $\mathcal{A}_{0s}$ and the local sieve space $\mathcal{A}_{0sn}$, we have

Theorem 3.1.1 Let $\gamma \equiv \min\{\gamma_1/2,\ \gamma_2/2,\ \gamma_3\} > 1/2$. Under assumptions 3.1.1–3.1.8, if $k_{1,n}^2 = O\left(n^{\frac{1}{\gamma_1+1}}\right)$, $k_{2,n}^2 = O\left(n^{\frac{1}{\gamma_2+1}}\right)$, and $k_{3,n} = O\left(n^{\frac{1}{2\gamma_3+1}}\right)$, then

$$\|\widehat{\alpha}_n - \alpha_0\|_2 = O_P\left(n^{\frac{-\gamma}{2\gamma+1}}\right) = o_P\left(n^{-1/4}\right).$$

Asymptotic Normality of Parametric Part

This section presents sufficient conditions for the asymptotic normality of the parametric part of the model. Define an inner product corresponding to the pseudo metric $\|\cdot\|_2$:

$$\langle v_1, v_2\rangle_2 \equiv -E\left[\frac{d^2\ell(D_i;\alpha_0)}{d\alpha d\alpha^T}[v_1, v_2]\right],$$

where

$$\frac{d^2\ell(D_i;\alpha_0)}{d\alpha d\alpha^T}[v_1,v_2] \equiv \frac{d^2\ell(D_i;\alpha_0+\tau_1 v_1+\tau_2 v_2)}{d\tau_1 d\tau_2}|_{\tau_1=\tau_2=0}.$$

Let $\overline{\mathbf{V}}$ denote the closure of the linear span of $\mathcal{A}-\{\alpha_0\}$ under the metric $\|\cdot\|_2$. Then $\left(\overline{\mathbf{V}},\|\cdot\|_2\right)$ is a Hilbert space. We define $\overline{\mathbf{V}}=\mathbb{R}^{d_\theta}\times\overline{\mathcal{U}}$ with $\overline{\mathcal{U}}\equiv\overline{\mathcal{F}_1\times\mathcal{F}_2\times\mathcal{F}_3}-\{(f_{01},f_{02},f_{03})\}$ and let $h=(f_1,f_2,f_3)$ denote all the unknown densities. The pathwise first derivative can be written as

$$\begin{aligned}\frac{d\ell(D_i;\alpha_0)}{d\alpha}[\alpha-\alpha_0] &= \frac{d\ell(D_i;\alpha_0)}{d\theta^T}(\theta-\theta_0)+\frac{d\ell(D_i;\alpha_0)}{dh}[h-h_0]\\ &= \left(\frac{d\ell(D_i;\alpha_0)}{d\theta^T}-\frac{d\ell(D;\alpha_0)}{dh}[\mu]\right)(\theta-\theta_0),\end{aligned}$$

with $h-h_0\equiv-\mu\times(\theta-\theta_0)$, and in which

$$\begin{aligned}\frac{d\ell(D_i;\alpha_0)}{dh}[h-h_0] &= \frac{d\ell(D_i;\theta_0,h_0(1-\tau)+\tau h)}{d\tau}|_{\tau=0}\\ &= \frac{d\ell(D_i;\alpha_0)}{df_1}[f_1-f_{01}]+\frac{d\ell(D_i;\alpha_0)}{df_2}[f_2-f_{02}]\\ &\quad+\frac{d\ell(D_i;\alpha_0)}{df_3}[f_3-f_{03}].\end{aligned}$$

Note that

$$\begin{aligned}&E\left(\frac{d^2\ell(D_i;\alpha_0)}{d\alpha d\alpha^T}[\alpha-\alpha_0,\alpha-\alpha_0]\right)\\ =\;&(\theta-\theta_0)^T E\left(\frac{d^2\ell(D_i;\alpha_0)}{d\theta d\theta^T}-2\frac{d^2\ell(D_i;\alpha_0)}{d\theta dh^T}[\mu]+\frac{d^2\ell(D_i;\alpha_0)}{dhdh^T}[\mu,\mu]\right)(\theta-\theta_0),\end{aligned}$$

with $h-h_0\equiv-\mu\times(\theta-\theta_0)$, and in which

$$\frac{d^2\ell(D_i;\alpha_0)}{d\theta dh^T}[h-h_0]=\frac{d(\partial\ell(D_i;\theta_0,h_0(1-\tau)+\tau h)/\partial\theta)}{d\tau}|_{\tau=0},$$

$$\frac{d^2\ell(D_i;\alpha_0)}{dhdh^T}[h-h_0,h-h_0]=\frac{d^2\ell(D_i;\theta_0,h_0(1-\tau)+\tau h)}{d\tau^2}|_{\tau=0}.$$

For each component θ^k (of θ), $k=1,...,d_\theta$, suppose there exists a $\mu^{*k}\in\overline{\mathcal{U}}$ that solves:

$$\mu^{*k}:\inf_{\mu^k\in\overline{\mathcal{U}}}E\left\{-\left(\frac{\partial^2\ell(D_i;\alpha_0)}{\partial\theta^k\partial\theta^k}-2\frac{d^2\ell(D_i;\alpha_0)}{\partial\theta^k dh^T}[\mu^k]+\frac{d^2\ell(D_i;\alpha_0)}{dhdh^T}[\mu^k,\mu^k]\right)\right\}.$$

Denote $\mu^*=\left(\mu^{*1},\mu^{*2},...,\mu^{*d_\theta}\right)$ with each $\mu^{*k}\in\overline{\mathcal{U}}$, and

$$\frac{d\ell(D_i;\alpha_0)}{dh}[\mu^*]=\left(\frac{d\ell(D_i;\alpha_0)}{dh}\left[\mu^{*1}\right],...,\frac{d\ell(D_i;\alpha_0)}{dh}\left[\mu^{*d_\theta}\right]\right),$$

$$\frac{d^2\ell(D_i;\alpha_0)}{\partial\theta dh^T}[\mu^*] = \left(\frac{d^2\ell(D_i;\alpha_0)}{\partial\theta dh}[\mu^{*1}], ..., \frac{d^2\ell(D_i;\alpha_0)}{\partial\theta dh}[\mu^{*d_\theta}]\right),$$

$$\frac{d^2\ell(D_i;\alpha_0)}{dhdh^T}[\mu^*,\mu^*] = \begin{pmatrix} \frac{d^2\ell(D_i;\alpha_0)}{dhdh^T}[\mu^{*1},\mu^{*1}] & \cdots & \frac{d^2\ell(D_i;\alpha_0)}{dhdh^T}[\mu^{*1},\mu^{*d_\theta}] \\ \cdots & \cdots & \cdots \\ \frac{d^2\ell(D_i;\alpha_0)}{dhdh^T}[\mu^{*d_\theta},\mu^{*1}] & \cdots & \frac{d^2\ell(D_i;\alpha_0)}{dhdh^T}[\mu^{*d_\theta},\mu^{*d_\theta}] \end{pmatrix}.$$

We also define

$$V_* \equiv -E\left(\frac{\partial^2\ell(D_i;\alpha_0)}{\partial\theta\partial\theta^T} - 2\frac{d^2\ell(D_i;\alpha_0)}{\partial\theta dh^T}[\mu^*] + \frac{d^2\ell(D_i;\alpha_0)}{dhdh^T}[\mu^*,\mu^*]\right). \tag{3.4}$$

We then consider a linear functional of α, which is $\lambda^T\theta$ for any $\lambda \in \mathbb{R}^{d_\theta}$ with $\lambda \neq 0$. Since

$$\begin{aligned} & \sup_{\alpha-\alpha_0\neq 0} \frac{|\lambda^T(\theta-\theta_0)|^2}{||\alpha-\alpha_0||_2^2} \\ = & \sup_{\theta\neq\theta_0,\mu\neq 0} \frac{(\theta-\theta_0)^T\lambda\lambda^T(\theta-\theta_0)}{(\theta-\theta_0)^T E\left\{-\left(\frac{d^2\ell(D_i;\alpha_0)}{d\theta d\theta^T} - 2\frac{d^2\ell(D_i;\alpha_0)}{d\theta dh^T}[\mu] + \frac{d^2\ell(D_i;\alpha_0)}{dhdh^T}[\mu,\mu]\right)\right\}(\theta-\theta_0)} \\ = & \lambda^T(V_*)^{-1}\lambda, \end{aligned}$$

the functional $\lambda^T(\theta-\theta_0)$ is *bounded* if and only if the matrix V_* is nonsingular.

Suppose that V_* is nonsingular. For any fixed $\lambda \neq 0$, denote $v^* \equiv (v_\theta^*, v_h^*)$ with $v_\theta^* \equiv (V_*)^{-1}\lambda$ and $v_h^* \equiv -\mu^* \times v_\theta^*$. Then the Riesz representation theorem implies: $\lambda^T(\theta-\theta_0) = \langle v^*, \alpha-\alpha_0\rangle_2$ for all $\alpha \in \mathcal{A}$. We have:

$$\begin{aligned} \lambda^T(\widehat{\theta}_n - \theta_0) &= \langle v^*, \widehat{\alpha}_n - \alpha_0\rangle_2 \\ &= \frac{1}{n}\sum_{i=1}^n \frac{d\ell(D_i;\alpha_0)}{d\alpha}[v^*] + o_p\{n^{-1/2}\}. \end{aligned} \tag{3.5}$$

Denote $\mathcal{N}_0 = \{\alpha \in \mathcal{A}_{0s} : \|\alpha-\alpha_0\|_2 = o(n^{-1/4})\}$ and $\mathcal{N}_{0n} = \{\alpha \in \mathcal{A}_{0sn} : \|\alpha-\alpha_0\|_2 = o(n^{-1/4})\}$. We provide additional sufficient for asymptotic normality of sieve MLE $\widehat{\theta}_n$ as follows:

Assumption 3.1.9 *μ^* exists (i.e., $\mu^{*k} \in \overline{\mathcal{U}}$ for $k = 1, ..., d_\theta$), and V_* is positive-definite.*

[1]

Assumption 3.1.10 *There is a $v_n^* \in \mathcal{A}_n - \{\alpha_0\}$, such that $||v_n^* - v^*||_2 = o(1)$ and $\|v_n^* - v^*\|_2 \times \|\widehat{\alpha}_n - \alpha_0\|_2 = o_P(\frac{1}{\sqrt{n}})$.*

[1] This assumption is necessary for the root-n convergence rate, but it may not always hold. See Chen and Liao (2014), Chen and Pouzo (2015), and Hahn and Liao (2018) for examples of ill-posed inverse problems in which the finite dimensional functionals fail to be root-n estimable.

Assumption 3.1.11 *There is a random variable $U(D_i)$ with $E\{[U(D_i)]^2\} < \infty$ and a non-negative measurable function η with $\lim_{\delta\to 0}\eta(\delta) = 0$, such that, for all $\alpha \in \mathcal{N}_{0n}$,*

$$\sup_{\overline{\alpha}\in\mathcal{N}_0}\left|\frac{d^2\ell(D_i;\overline{\alpha})}{d\alpha d\alpha^T}[\alpha-\alpha_0,\upsilon_n^*]\right| \leq U(D_i)\times\eta(||\alpha-\alpha_0||_s).$$

Assumption 3.1.12 *Uniformly over $\overline{\alpha}\in\mathcal{N}_0$ and $\alpha\in\mathcal{N}_{0n}$,*

$$E\left(\frac{d^2\ell(D_i;\overline{\alpha})}{d\alpha d\alpha^T}[\alpha-\alpha_0,\upsilon_n^*] - \frac{d^2\ell(D_i;\alpha_0)}{d\alpha d\alpha^T}[\alpha-\alpha_0,\upsilon_n^*]\right) = o\left(\frac{1}{\sqrt{n}}\right).$$

Assumption 3.1.13 *$E\{\left(\frac{d\ell(D_i;\alpha_0)}{d\alpha}[\upsilon_n^*-\upsilon^*]\right)^2\}$ goes to zero as $\|\upsilon_n^*-\upsilon^*\|_2$ goes to zero.*

Recall the definitions of Fisher inner product and the Fisher norm:

$$\langle v_1,v_2\rangle \equiv E\left\{\left(\frac{d\ell(D_i;\alpha_0)}{d\alpha}[v_1]\right)\left(\frac{d\ell(D_i;\alpha_0)}{d\alpha}[v_2]\right)\right\}, \quad \|v\| \equiv \sqrt{\langle v,v\rangle}.$$

Under correct specification, $m(y^*;\theta_0) = E(z|y^*,l)$, it can be shown that $\|v\| = \|v\|_2$ and $\langle v_1,v_2\rangle = \langle v_1,v_2\rangle_2$. Thus, the space $\overline{\mathbf{V}}$ is also the closure of the linear span of $\mathcal{A}-\{\alpha_0\}$ under the Fisher metric $\|\cdot\|$.
Suppose that θ has d_θ components, and write its k^{th} component as θ^k. Write $\mu^* = \left(\mu^{*1},\mu^{*2},...,\mu^{*d_\theta}\right)$, where we compute $\mu^{*k}\equiv\left(\mu_1^{*k},\mu_2^{*k},\mu_3^{*k}\right)^T\in\overline{\mathcal{U}}$ as the solution to

$$\begin{aligned}&\inf_{\mu^k\in\overline{\mathcal{U}}} E\left\{\left(\frac{d\ell(D_i;\alpha_0)}{d\theta^k} - \frac{d\ell(D_i;\alpha_0)}{dh}\left[\mu^k\right]\right)^2\right\}\\ &\quad= \inf_{(\mu_1,\mu_2,\mu_3)^T\in\overline{\mathcal{U}}} E\left\{\begin{pmatrix}\frac{d\ell(D_i;\alpha_0)}{d\theta^k} - \frac{d\ell(D_i;\alpha_0)}{df_1}[\mu_1]\\ -\frac{d\ell(D_i;\alpha_0)}{df_2}[\mu_2] - \frac{d\ell(D_i;\alpha_0)}{df_3}[\mu_3]\end{pmatrix}^2\right\}.\end{aligned}$$

This equation also defines $\frac{d\ell(D_i;\alpha_0)}{dh}[\mu^*]$. Then $\mathcal{S}_{\theta_0}\equiv\frac{d\ell(D_i;\alpha_0)}{d\theta^T} - \frac{d\ell(D_i;\alpha_0)}{dh}[\mu^*]$ becomes the semiparametric efficient score for θ_0, and

$$I_* \equiv E\left[\mathcal{S}_{\theta_0}^T\mathcal{S}_{\theta_0}\right] = V_* \tag{3.6}$$

becomes the semiparametric information bound for θ_0.

Finally, we can show that the sieve MLE $\widehat{\theta}_n$ is asymptotically normally distributed around θ_0 as follows:

Theorem 3.1.2 *Suppose that Assumptions of Lemma 3.1.1, and Assumptions 3.1.5–3.1.13 hold. Then: $\sqrt{n}(\widehat{\theta}_n-\theta_0)\xrightarrow{d}N(0,V_*^{-1}I_*V_*^{-1})$, with V_* defined in equation (3.4) and I_* given by equation (3.6).*

3.2 Closed-form Estimators

In most of the existing estimators, there exists a gap between identification and estimation in the sense that not all the identification conditions are imposed in the estimation procedure. This section introduces a class of estimators, which may fill this gap at the expense of efficiency.

Although the sieve MLE in (3.3) is quite general and flexible, a few identification results in this section provide closed-form expressions for the unobserved components as functions of observed distribution functions, which can lead to straightforward closed-form estimators. In the case where X^* is continuous, for example, Li and Vuong (1998) suggest that the distribution of the latent variable f_{X^*} in equation (2.31) can be estimated using Kotlarski's identity with characteristic functions being replaced by corresponding empirical characteristic functions. In general, one can consider a nonlinear regression model in the framework of the 3-measurement model as

$$\begin{aligned} Y &= g_1(X^*) + \eta \\ X &= g_2(X^*) + \epsilon \\ Z &= g_3(X^*) + \epsilon' \end{aligned} \tag{3.7}$$

where ϵ and ϵ' are independent of X^* and η with $E[\eta|X^*] = 0$. Since X^* is unobserved, we may normalize $g_3(X^*) = X^*$. Schennach (2004b) provides a closed-form estimator of $g_1(\cdot)$ in the case where $g_2(X^*) = X^*$ using Kotlarski's identity as follows:[2]

$$g_1(x^*) = \frac{\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-it_1x^*}\left(\frac{\left[\frac{\partial}{\partial s}\phi_{X,Y}(t_1,s)\right]_{s=0}}{i\phi_X(t_1)}\phi_{X^*}(t_1)\right)dt_1}{\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx^*}\phi_{X^*}(t)\,dt}$$

where f_{X^*} is identified from the characteristic function

$$\phi_{X^*}(t) = \exp\left(\int_0^t \frac{\left[\frac{\partial}{\partial t_2}\phi_{X,Z}(s,t_2)\right]_{t_2=0}}{i\phi_X(s)}ds\right).$$

Hu and Sasaki (2015) generalize that estimator to the case where $g_2(\cdot)$ is a polynomial. Whether a closed-form estimator of $g_1(\cdot)$ exists or not with a general $g_2(\cdot)$ is a challenging and open question for future research.

In the case where X^* is discrete as in Theorem 2.4.1 and Corollary 2.5.1, the sieve MLE is still applicable. Nevertheless, the identification strategy in the discrete case also leads to a closed-form estimator for the unknown probabilities in the sense that one can mimic the identification procedure to solve for the unknowns. In estimation, it is more convenient to

[2]Schennach (2007) also provides a closed-form estimator for a similar nonparametric regression model using a generalized function approach.

use the equation below than directly using Equation (2.36)

$$E\left[\omega\left(Y\right)|X=x,Z=z\right]f_{X,Z}\left(x,z\right)=\sum_{x^{*}\in\mathcal{X}^{*}}f_{X|X^{*}}(x|x^{*})E\left[\omega\left(Y\right)|x^{*}\right]f_{Z|X^{*}}(z|x^{*})f_{X^{*}}(x^{*}), \tag{3.8}$$

which leads to an eigenvalue-eigenvector decomposition

$$M_{X,\omega,Z}M_{X,Z}^{-1}=M_{X|X^{*}}D_{\omega|X^{*}}M_{X|X^{*}}^{-1} \tag{3.9}$$

with

$$\begin{aligned} M_{X,\omega,Z} &= \left[E\left[\omega\left(Y\right)|X=x_{k},Z=z_{l}\right]f_{X,Z}\left(x_{k},z_{l}\right)\right]_{k=1,2,...,K;l=1,2,...,K} \\ D_{\omega|X^{*}} &= diag\left\{E\left[\omega\left(Y\right)|x_{1}^{*}\right],E\left[\omega\left(Y\right)|x_{2}^{*}\right],\ldots,E\left[\omega\left(Y\right)|x_{K}^{*}\right]\right\}. \end{aligned} \tag{3.10}$$

The matrix $M_{X,\omega,Z}$ can be directly estimated as

$$\widehat{M_{X,\omega,Z}}=\left[\frac{1}{N}\sum_{i=1}^{N}\omega\left(Y_{i}\right)\mathbf{1}\left(X_{i}=x_{k},Z_{i}=z_{l}\right)\right]_{k=1,2,...,K;l=1,2,...,K}$$

where $\mathbf{1}(\cdot)$ is the indicator function. Similarly, matrix $M_{X,Z}$ can be estimated as $\widehat{M_{X,Z}}=\widehat{M_{X,\omega,Z}}\Big|_{\omega(\cdot)=1}$. Solving for eigenvectors and eigenvalues in Equation (3.9) can be considered as a procedure to minimize the Euclidean distance $\|\cdot\|$ between the left hand side and the right hand side of that equation, in fact, to zero. Moreover, Assumption 2.4.4 can be directly used to order the eigenvectors or the eigenvalues. With a finite sample, estimated probabilities might be outside $[0,1]$ or even a complex number. One remedy is to use Equation (3.9) as a moment condition to estimate the unknown probabilities under suitable restrictions. To be specific, matrices $M_{X|X^{*}}$ and $D_{\omega|X^{*}}$ can be estimated as follows:

$$\begin{aligned} \left(\widehat{M_{X|X^{*}}},\widehat{D_{\omega|X^{*}}}\right) &= \arg\min_{M,D}\left\|\widehat{M_{X,\omega,Z}}\left(\widehat{M_{X,Z}}\right)^{-1}M-M\times D\right\| \\ &\quad \text{such that} \\ &\quad 1)\text{ each entry in } M \text{ is in } [0,1]\ ; \\ &\quad 2)\text{ each column sum of } M \text{ equals 1 and } D \text{ is diagonal;} \\ &\quad 3)\text{ entries in } M \text{ and } D \text{ satisfy Assumptions 2.4.3 and 2.4.4.} \end{aligned} \tag{3.11}$$

This closed-form estimator performs well in empirical studies, such as An et al. (2017) , An et al. (2010) , Feng and Hu (2013) , and Hu et al. (2013b) .

Such closed-form estimators may not be as efficient as the sieve MLE, but they have their advantages that there are much fewer nuisance parameters involved than indirect estimators and that the computation of closed-form estimators may not rely on optimization algorithms, which usually involve many iterations and are time-consuming. An optimization algorithm can only guarantee a local maximum or minimum, while a closed-form estimator is a global one by construction. Although a closed-form estimator may not always exist, it is much more straightforward and transparent, if available, than an indirect estimator.

Such closed-form estimation may be a challenging but useful approach for future research.

3.2.1 Regression with Misclassification: Simulations and Code

This section provides simulation results for both the minimum distance estimator and the plug-in estimator proposed for a linear regression model with a misclassified regressor and its two measurements. [3] We consider a linear regression model of a wage equation as follows:

$$Y = m(X^*) + \eta$$

where Y is log wage, X^* is the true education level, and η is a regression error with a normal distribution $N(0, 0.5^2)$. Instead of observing X^*, we observed two measurements, i.e., the self-reported education attainment X and the transcript-recorded education attainment Z

We generate the data as follows: First, we generate the true education attainment X^* using its marginal distribution; Second, we generate the self-reported education attainment X and the transcript-recorded education attainment Z using their associated error matrices separately. We also generate the wage corresponding to the true education attainment with the normal distribution as the wage shocks. We replicate the above process for 1000 times for sample size (N) 500 and 1000, respectively.

As discussed above, we consider two estimators, i.e., a minimum distance estimator in Equation (3.11) and a plug-in estimator following step-by-step the identification procedure using the eigenvalue-eigenvector decomposition in Equation (3.9). Tables 3.1 and 3.2 provide the true values and the estimates of the mean wages and the marginal probabilities of the latent education levels. Table 3.3 presents the true values of the misclassification probability in the two measurements. Tables 3.4, 3.5, 3.6, and 3.7 include the estimation results in the simulations. Click here ↗ to download the Matlab code for the simulations.

Table 3.1: Estimates of Mean Wage

	True	Min-distance		Plug-in Estimator	
		N=500	N=1000	N=500	N=1000
High school	2.0250	2.0227***	2.0249***	2.0232***	2.0252***
		(0.0411)	(0.0289)	(0.0419)	(0.0293)
Some College	2.2074	2.2081***	2.2073***	2.2073***	2.2071***
		(0.0446)	(0.0312)	(0.0461)	(0.0317)
Bachelor	2.4456	2.4451***	2.4441***	2.4446***	2.4438***
		(0.0391)	(0.0275)	(0.0394)	(0.0278)

[1] Standard errors (calculated by 1000 replications) are in parentheses.
[2] Symbols *** indicate that the test is significant at a level of 1%.

[3] I am grateful for Professor Ruli Xiao at Indiana University for providing MATLAB codes and other materials in this and next sections.

Table 3.2: Estimates of True Education Attainment's Marginal Probabilities

	True	Min-distance		Plug-in Estimator	
		N=500	N=1000	N=500	N=1000
High school	0.3432	0.3412***	0.3413***	0.3432***	0.3424***
		(0.0393)	(0.0328)	(0.0623)	(0.0376)
Some College	0.3022	0.3073***	0.3060***	0.3084***	0.3056***
		(0.0381)	(0.0329)	(0.0637)	(0.0390)
Bachelor	0.3546	0.3515***	0.3526***	0.3484***	0.3521***
		(0.0236)	(0.0181)	(0.0341)	(0.0188)

[1] Standard errors (calculated by 1000 replications) are in parentheses.
[2] Symbols *** indicate that the test is significant at a level of 1%.

Table 3.3: DGP: True Misclassification Probability for both Measurements

		Conditional on true education level:		
		High School	Some College	Bachelor
Self-Reported	High School	0.8838	0.0630	0.0001
	Some College	0.1049	0.9111	0.0095
	Bachelor	0.0113	0.0259	0.9904
Transcript	High School	0.9439	0.0928	0.0291
	Some College	0.0557	0.9007	0.0494
	Bachelor	0.0004	0.0065	0.9215

Table 3.4: Estimated Error Probability (Min-distance), N=500

		Conditional on true education level:		
		High School	Some College	Bachelor
Self-reported	High school	0.8881	0.0631	0.0004
		(0.0778)	(0.0630)	(0.0015)
	Some College	0.0977	0.9051	0.0104
		(0.0757)	(0.0679)	(0.0095)
	Bachelor	0.0142	0.0318	0.9892
		(0.0154)	(0.0345)	(0.0096)
Transcript-recorded	High school	0.9453	0.0947	0.0254
		(0.0625)	(0.0845)	(0.0172)
	Some College	0.0544	0.9005	0.0447
		(0.0624)	(0.0842)	(0.0292)
	Bachelor	0.0004	0.0048	0.9299**
		(0.0014)	(0.0082)	(0.0344)

[1] Standard errors (calculated by 1000 replications) are in parentheses.
[2] We separately test the diagonal element being 1 and the off-diagonal element being zero. Symbols ***, **, and * indicate that the test is significant at a level of 1%, 5%, and 10%, respectively.

Table 3.5: Estimated Error Probability (Min-distance), N=1000

		Conditional on true education level:		
		High School	Some College	Bachelor
Self-reported	High school	0.8893**	0.0625	0.0004
		(0.0570)	(0.0450)	(0.0011)
	Some College	0.0981*	0.9069*	0.0107
		(0.0552)	(0.0524)	(0.0076)
	Bachelor	0.0126	0.0306	0.9889
		(0.0116)	(0.0289)	(0.0076)
Transcript-recorded	High school	0.9466	0.0920	0.0273**
		(0.0404)	(0.0634)	(0.0132)
	Some College	0.0531	0.9033	0.0450*
		(0.0403)	(0.0631)	(0.0251)
	Bachelor	0.0003	0.0047	0.9277***
		(0.0008)	(0.0069)	(0.0296)

[1] Standard errors (calculated by 1000 replications) are in parentheses.

[2] We separately test the diagonal element being 1 and the off-diagonal element being zero. Symbols ***, **, and * indicate that the test is significant at a level of 1%, 5%, and 10%, respectively.

Table 3.6: Estimated Error Probability (Plug-in), N=500

		Conditional on true education level:		
		High School	Some College	Bachelor
Self-reported	High school	0.8761	0.0727	0.0008
		(0.0867)	(0.0746)	(0.0026)
	Some College	0.1083	0.8895	0.0130
		(0.0850)	(0.0824)	(0.0180)
	Bachelor	0.0157	0.0378	0.9862
		(0.0177)	(0.0452)	(0.0185)
Transcript-recorded	High school	0.9314	0.1028	0.0282
		(0.0759)	(0.0996)	(0.0207)
	Some College	0.0676	0.8834	0.0525
		(0.0758)	(0.0991)	(0.0408)
	Bachelor	0.0010	0.0138	0.9193*
		(0.0027)	(0.0196)	(0.0452)

[1] Standard errors (calculated by 1000 replications) are in parentheses.

[2] We separately test the diagonal element being 1 and the off-diagonal element being zero. Symbols ***, **, and * indicate that the test is significant at a level of 1%, 5%, and 10%, respectively.

Table 3.7: Estimated Error Probability (Plug-in), N=1000

		Conditional on true education level:		
		High School	Some College	Bachelor
Self-reported	High school	0.8833*	0.0661	0.0007
		(0.0632)	(0.0513)	(0.0017)
	Some College	0.1034*	0.9013*	0.0117
		(0.0623)	(0.0596)	(0.0126)
	Bachelor	0.0133	0.0325	0.9876
		(0.0122)	(0.0321)	(0.0129)
Transcript-recorded	High school	0.9401	0.0932	0.0287**
		(0.0470)	(0.0672)	(0.0147)
	Some College	0.0590	0.8962	0.0495
		(0.0471)	(0.0680)	(0.0324)
	Bachelor	0.0009	0.0105	0.9219**
		(0.0020)	(0.0135)	(0.0362)

[1] Standard errors (calculated by 1000 replications) are in parentheses.

[2] We separately test the diagonal element being 1 and the off-diagonal element being zero. Symbols ***, **, and * indicate that the test is significant at a level of 1%, 5%, and 10%, respectively.

3.2.2 Misclassification in Education: Data, Code, and Estimates

This section provides an empirical example for methods proposed above with a Matlab code using the dataset in Kane et al. (1999), which contains wages and education levels from the National Longitudinal Study of 1972 and the Postsecondary Education Transcript Study. Table 3.8 provides the summary statistics of the data. Tables 3.9 and 3.11 provide the estimates of the error probability matrix for both transcript-recorded and self-reported education attainments. In particular, Table 3.9 uses a minimum distance estimator in Equation (3.11); and Table 3.11 uses the plug-in estimator following step-by-step the identification procedure using the eigenvalue-eigenvector decomposition in Equation (3.9). The minimum distance estimator requires an initial estimate for optimization iterations, which is usually ignored in the discussion of estimation. Given the advantages of the closed-form estimator, a natural choice of such an initial estimate is the plug-in estimates in Table 3.11. The ordering assumption is imposed on the eigenvalues such that average wages increase with education levels. The standard errors are computed through bootstrap. Click here ↗ to download the datasets and the Matlab code, which runs about 26 minutes.

Table 3.8: Sample Proportions and Mean Log Wages in 1986

Sample Proportions:				
	Self-Reported Schooling:			
Transcript-recorded	High School	Some College	Bachelor's Degree	Row Total
High School	0.2881	0.0596	0.0146	0.3623
Some College	0.0340	0.2502	0.0246	0.3088
Bachelor's Degree	0.0200	0.5000	0.3237	0.3289
Column Total	0.3223	0.33148	0.3629	1.0000
Mean Log Wages in 1986:				
	Self-Reported Schooling:			
Transcript-recorded	High School	Some College	Bachelor's Degree	Row Total
High School	2.0262	2.1039	2.3304	2.0512
	(0.4970)	(0.4941)	(0.6094)	(0.5055)
Some College	2.1175	2.2061	2.3748	2.2098
	(0.4534)	(0.4882)	(0.5162)	(0.4916)
Bachelor's Degree	2.4604	2.3101	2.4456	2.4435
	(0.3503)	(0.4166)	(0.4947)	(0.4937)
Column Total	2.0360	2.1884	2.4362	2.2292
	(0.4950)	(0.4900)	(0.5018)	(0.5235)

[1] Educational attainment measured as of 1979, and average log hourly wages observed in 1986. The sample size is 9261.

[2] Source: NLS-72 and PETS.

Table 3.9: Estimated Error Probability for both Measures (a Minimal Distance Estimator)

		Conditional on true education level:		
		High School	Some College	Bachelor
Self-Reported	High School	0.8839***	0.0630***	0.0001
		(0.0220)	(0.0189)	(0.0001)
	Some College	0.1049***	0.9112***	0.0095***
		(0.0209)	(0.0218)	(0.0037)
	Bachelor	0.0113**	0.0259**	0.9905***
		(0.0056)	(0.0128)	(0.0037)
Transcript	High School	0.9440***	0.0928***	0.0291***
		(0.0172)	(0.0251)	(0.0054)
	Some College	0.0557***	0.9008***	0.0494***
		(0.0171)	(0.0252)	(0.0101)
	Bachelor	0.0004	0.0065	0.9215***
		(0.0004)	(0.0044)	(0.0116)

[1] Standard errors (calculated by bootstrap) are in parentheses.

[2] We separately test the diagonal element being 1 and the off-diagonal element being zero. Symbols ***, **, and * indicate that the test is significant at a level of 1%, 5%, and 10%, respectively.

Table 3.10: Mean wage and marginal distribution of true education

		High School	Some College	Bachelor
Min-distance estimator	Mean wage	2.0250***	2.2074***	2.4456***
		(0.0097)	(0.0106)	(0.0087)
	Marginal distribution	0.3432***	0.3022***	0.3546***
		(0.0127)	(0.0129)	(0.0066)
Plug-in estimator	Mean wage	2.0250***	2.2075***	2.4456***
		(0.0097)	(0.0106)	(0.0087)
	Marginal distribution	0.3427***	0.3035***	0.3538***
		(0.0125)	(0.0127)	(0.0065)

[1] Standard errors (calculated by bootstrap) are in parentheses.
[2] Symbols *** indicate that the test is significant at a level of 1%.

Table 3.11: Estimated Error Probability for both Measures (a Plug-in Estimator)

		Conditional on true education level:		
		High School	Some College	Bachelor
Self-Reported	High School	0.8842***	0.0632***	0.0004
		(0.0220)	(0.0184)	(0.0005)
	Some College	0.1047***	0.9113***	0.0065*
		(0.0210)	(0.0212)	(0.0036)
	Bachelor	0.0111**	0.0255**	0.9930*
		(0.0055)	(0.0127)	(0.0037)
Transcript	High School	0.9446***	0.0926***	0.0292***
		(0.0165)	(0.0250)	(0.0054)
	Some College	0.0554***	0.8970***	0.0497***
		(0.0165)	(0.0250)	(0.0102)
	Bachelor	0.000	0.0103**	0.9211***
		(0.0001)	(0.0047)	(0.0117)

[1] Standard errors (calculated by bootstrap) are in parentheses.
[2] We separately test the diagonal element being 1 and the off-diagonal element being zero. Symbols ***, **, and * indicate that the test is significant at a level of 1%, 5%, and 10%, respectively.

3.2.3 Regressions with Non-Classical Measurement Errors

Here we present the closed-form estimation of nonparametric regression models with non-classical measurement errors in Hu and Sasaki (2015). Specifically, we explicitly estimate the nonparametric regression function g for the model

$$Y = g(X^*) + U \qquad E\left[U|X^*\right] = 0,$$

where Y is an observed dependent variable, X^* is an unobserved explanatory variable, and U is the regression residual. While the true explanatory variable X^* is not observed, two measurements, X_1 and X_2, are available from matched data. For simplicity, X^* is assumed

to be a scalar and continuously distributed. The relationship between the two measurements and the true explanatory variable X^* is modeled as follows.

$$\begin{aligned} X_1 &= \sum_{p=0}^{P} \gamma_p X^{*p} + \epsilon_1 \\ X_2 &= X^* + \epsilon_2 \end{aligned}$$

Unless $\gamma_0 = 0$, $\gamma_1 = 1$ and $\gamma_2 = \cdots = \gamma_P = 0$ are true, the first measurement X_1 entails non-classical errors with nonlinearity. Allowing for such non-classical errors is crucial particularly for survey data that are often contaminated by endogenous self-reporting biases. Since the truth X^* is unobserved, the second measurement X_2 is location-/scale-normalized with respect to the unobserved truth X^*. We use alternative independence assumptions on the measurement error ϵ_2 depending on which order P we assume about X_1, but these assumptions are more innocuous than assuming classical errors in any case.

Under assumptions that will be introduced below, we show that the regression function g can be explicitly expressed as a functional of the joint CDF $F_{YX_1X_2}$ in the following sense.

$$g(x^*) = \lambda(x^* | F_{YX_1X_2}).$$

We provide the concrete expression for this functional $\lambda(x^* \mid \cdot\ )$. In order to construct a sample-counterpart estimator of $g(x^*)$ given this closed-form identifying solution, it suffices to substitute the empirical distribution $\widehat{F}_{YX_1X_2}$ in this known transformation so we get the closed-form estimator $\widehat{g(x^*)} = \lambda(x^* \mid \widehat{F}_{YX_1X_2})$. Measurement error models have been extensively studied in both statistics and econometrics. The statistical literature focuses on cases of classical errors, where measurement errors are independent of the true values – see Fuller (1987) and Carroll, Ruppert, Stefanski and Crainiceanu (2006) for reviews. The econometric literature investigates nonlinear models and nonclassical measurement errors – see Chen, Hong and Nekipelov (2011), Bound, Brown and Mathiowetz (2001) and Schennach (2013) for reviews. However, closed-form estimation, nonlinear/nonparametric models, and non-classical measurement errors still remain unsolved, despite their joint practical relevance. Two measurements are known to be useful to correct measurement errors even for external samples if the matched administrative data is known to be true (e.g., Chen, Hong, and Tamer, 2005). The baseline model of our framework was introduced by Li (2002) and Schennach (2004a), where they consider parametric regression models under two measurements with classical errors. Hu and Schennach (2008) provide general identification results for nonseparable and non-classical measurement errors,[4] but their estimator relies on semi-/non-parametric extremal estimator where nuisance functions are approximated by truncated series. [5] Unlike these existing approaches, we develop a closed-form estimator

[4] Also see Mahajan (2006), Lewbel (2007), and Hu (2008) for non-/semi-parametric identification and estimation under non-classical measurement errors with discrete variables.

[5] Our model is also closely related to nonparametric regression models with classical measurement errors, which are extensively studied in the rich literature in statistics. When the error distribution is known, the regression function may be estimated by deconvolution – see Fan and Truong (1993) and Carroll, Ruppert, Stefanski and Crainiceanu (2006) for reviews. When the error distribution is unknown, Schennach (2004b)

for nonparametric models involving non-classical measurement errors.

Our results share much in common with Schennach (2004b) where she develops a closed-form estimator under the restriction, $\gamma_1 = 1$ and $\gamma_2 = \cdots = \gamma_P = 0$, of a classical-error structure. There are notable differences and thus values added by this paper as well. Our method paves the way for non-classical error structures with high degrees of nonlinearity whereas the existing closed-form estimator can handle only classical errors. To this end, we propose a new method to recover and use the characteristic function of the generated latent variable $\sum_{p=1}^{P} \gamma_p X^{*p}$, instead of just X^*, in the framework of deconvolution approaches. Not surprisingly, as we show through simulations, the classical error assumption $\gamma_1 = 1$ and $\gamma_2 = \cdots = \gamma_P = 0$ can severely bias estimates if the true DGP does not conform with this assumption. In our empirical application, we find that $\gamma_1 \neq 1$ is indeed true when people report their physical characteristics, and hence the existing closed-form estimator that assumes classical errors would likely suffer from biased estimates. The contribution of our method is to overcome these practical limitations of the existing closed-form estimators.

Closed-Form Identification: A Baseline Model

Our objective is to derive closed-form identifying formulas for the nonparametric regression function g. For the purpose of intuitive exposition, we first focus on the following simple model:

$$\begin{aligned} Y &= g(X^*) + U, & E[U \mid X^*] = 0 \\ X_1 &= \gamma_1 X^* + \epsilon_1 & E[\epsilon_1] = \gamma_0 \\ X_2 &= X^* + \epsilon_2, & E[\epsilon_2] = 0 \end{aligned} \tag{3.12}$$

where we observe the joint distribution of (Y, X_1, X_2). The restriction $E[U \mid X^*] = 0$ means that $g(X^*)$ is the nonparametric regression of Y on X^*. We do not assume $E[\epsilon_1]$ to be zero in order to accommodate arbitrary intercept γ_0 for the first measurement X_1. As such, we suppress γ_0 from the equation for X_1, i.e., it is embedded in $\gamma_0 = E[\epsilon_1]$. On the other hand, the locational normalization $E[\epsilon_2] = 0$ is imposed on the second measurement X_2. A leading example of (3.12) is the case with $\gamma_1 = 1$ often assumed in related papers in the literature. We do not make such an assumption, and thus our model (3.12) accommodates the possibility that the first measurement X_1 is endogenously biased even if $X^* \perp \epsilon_1$ is assumed, as $E[X_1 - X^* \mid X^*] = \gamma_0 + (\gamma_1 - 1)X^*$.

We can easily show that γ_1 is identified from the observed data by the closed-form formula:

$$\gamma_1 = \frac{\mathrm{Cov}(Y, X_1)}{\mathrm{Cov}(Y, X_2)} \tag{3.13}$$

under the following assumption.

Assumption 3.2.1 (Identification of γ_1) *$Cov(\epsilon_1, Y) = Cov(\epsilon_2, Y) = 0$ and $Cov(Y, X_2) \neq 0$.*

uses Kotlarski's identify (see Rao, 1992) to provide a Nadaraya-Watson-type estimator for the regression function.

The first part of this assumption requires that ϵ_1 and ϵ_2 are uncorrelated with the dependent variable. These zero covariance restrictions can be implied by a lower-level assumption, such as $E[U \mid X^*, \epsilon_1, \epsilon_2] = 0$, $\epsilon_1 \perp X^*$, and $E[\epsilon_2 \mid X^*] = 0$, which also imply the additional identifying restrictions presented later (Assumption 3.2.3). The second part of Assumption 3.2.1 is empirically testable with observed data, and implies a non-zero denominator in the identifying equation (3.13). We state this auxiliary result below for ease of reference.

Lemma 3.2.1 (Identification of γ_1) *If Assumption 3.2.1 holds, then γ_1 is identified with (3.13).*

In some applications, we may simply assume $\gamma_1 = 1$ from the outset, and Assumption 3.2.1 need not be invoked. In any case, we hereafter assume that γ_1 is known either by assumption or by the identifying formula (3.13), and that γ_1 is different from zero.

Assumption 3.2.2 (Nonzero γ_1) $\gamma_1 \neq 0$.

If this assumption fails, then the observed variable X_1 fails to be an informative signal of X^*. Assumption 3.2.2 therefore plays the role of letting X_1 be an effective proxy for the latent variable X^*. To complete our definition of the model (3.12), we impose the following independence restrictions.

Assumption 3.2.3 (Identifying Restrictions)
(i) $E[U|X_1] = 0$, *(ii)* $\epsilon_1 \perp X^*$, *and (iii)* $E[\epsilon_2|X_1] = 0$

Part (i) states that the residual of the outcome equation is conditional mean independent of the first measurement. A stronger version of part (i) is the mean independence $E[U|X^*, \epsilon_1] = 0$. Part (ii) states that the random error ϵ_1 in X_1 is independent of the true explanatory variable X^*. Notice that the coefficient γ_1 may not equal to one, and therefore the first measurement error defined as $X_1 - X^* = (\gamma_1 - 1)X^* + \epsilon_1$ need not be classical, i.e., the measurement error is not independent of the true value X^*, even under part (ii) of the above assumption. This observation highlights one of the major advantages of our model compared to the existing models which impose $\gamma_1 = 1$. Part (iii) states that the second measurement error ϵ_2 is conditional mean independent of the first measurement X_1. This assumption is different from the classical measurement error assumption that ϵ_2 is independent of X^* and U. The last two parts, (ii) and (iii), can be succinctly implied by the frequently used assumption in the literature that X^*, ϵ_1, and ϵ_2 are mutually independent, but we state the above weaker assumptions for the sake of generality.

Let $i = \sqrt{-1}$ denote the unit imaginary number. Define the marginal characteristic functions ϕ_{X_1}, ϕ_{X^*} and ϕ_{ϵ_1} by

$$\phi_{X_1}(t) = Ee^{itX_1}, \qquad \phi_{X^*}(t) = Ee^{itX^*} \qquad \text{and} \qquad \phi_{\epsilon_1}(t) = Ee^{it\epsilon_1},$$

respectively. Also define the joint characteristic functions $\phi_{X_1X_2}$ and ϕ_{X_1Y} by

$$\phi_{X_1X_2}(t_1, t_2) = Ee^{it_1X_1 + it_2X_2} \qquad \text{and} \qquad \phi_{X_1Y}(t_1, s) = Ee^{it_1X_1 + isY},$$

respectively. We let $\mathcal{F}$ denote the transformation defined by

$$\mathcal{F}f(t) \;=\; \int e^{itx} f(x)dx.$$

With this notation, we state the following assumption for identification of g.

Assumption 3.2.4 (Regularity) *(i) ϕ_{X_1} does not vanish on the real line. (ii) f_{X^*} and $\mathcal{F}f_{X^*}$ are continuous and absolutely integrable. (iii) $f_{X^*} \cdot g$ and $\mathcal{F}(f_{X^*} \cdot g)$ are continuous and absolutely integrable.*

Under Assumptions 3.2.3 (ii) and 3.2.4 (i), the characteristic functions ϕ_{X^*} and ϕ_{ϵ_1} do not vanish on the real line either. This property of non-vanishing characteristic functions is shared by many of the common distribution families, e.g., the normal, chi-squared, Cauchy, gamma, and exponential distributions. In parts of our identifying formula, the characteristic functions appear as denominators, and hence this assumption to rule out zero denominator is crucial. Parts (ii) and (iii) ensure that we can apply the Fourier transform and inversion to those functions. Under this commonly invoked regularity condition together with the independence restrictions in Assumption 3.2.3, we can solve relevant integral equations explicitly to obtain the following closed-form identification result.

Theorem 3.2.1 *Suppose that Assumptions 3.2.1, 3.2.2, 3.2.3 and 3.2.4 hold for the model (3.12). The nonparametric function g evaluated at x^* in the interior of the support of X^* is identified with the closed-form solution:*

$$g(x^*) = \frac{\int_{-\infty}^{+\infty} e^{-itx^*} \;\exp\left(\int_0^t \frac{\left[\frac{\partial}{\partial t_2}\phi_{X_1X_2}(t_1/\gamma_1,t_2)\right]_{t_2=0}}{\phi_{X_1}(t_1/\gamma_1)}dt_1\right) \;\frac{\left[\frac{\partial}{\partial s}\phi_{X_1Y}(t/\gamma_1,s)\right]_{s=0}}{i\;\phi_{X_1}(t/\gamma_1)}\;dt}{\int_{-\infty}^{+\infty} e^{-itx^*}\exp\left(\int_0^t \frac{\left[\frac{\partial}{\partial t_2}\phi_{X_1X_2}(t_1/\gamma_1,t_2)\right]_{t_2=0}}{\phi_{X_1}(t_1/\gamma_1)}dt_1\right)dt}, \tag{3.14}$$

where the parameter γ_1 is identified with the closed-form solution (3.13).

Note that every component on the right-hand side of the identifying formula (3.14) is computable directly as a moment of observed data. Replacing the population moments by the corresponding sample moments therefore yields a closed-form estimator of $g(x^*)$.

Closed-Form Identification: General Models

In this section, we consider the following generalized extension to the baseline model (3.12):

$$\begin{aligned} Y &= g(X^*) + U, & E[U \mid X^*] = 0 \\ X_1 &= \sum_{p=1}^{P} \gamma_p X^{*p} + \epsilon_1 & E[\epsilon_1] = \gamma_0 \\ X_2 &= X^* + \epsilon_2, & E[\epsilon_2] = 0 \end{aligned} \tag{3.15}$$

where we observe the joint distribution of (Y, X_1, X_2). The first measurement X_1 is systematically biased with an arbitrarily high order of nonlinearity. We demonstrate that a similar closed-form identification result can be obtained for this extended model. To this goal, we impose the following independence restrictions on (3.15).

Assumption 3.2.5 (Identifying Restrictions for the General Polynomial Model)

(i) $E[U \mid X^*, \epsilon_1, \epsilon_2] = 0$, *(ii)* $X^* \perp \epsilon_1$, *and (iii)* $(X^*, \epsilon_1) \perp \epsilon_2$.

Parts (i)–(iii) of this assumption are analogous to the corresponding parts in Assumption 3.2.3. We remark that parts (i) and (iii) are stronger than those corresponding parts in Assumption 3.2.3, and that we can deal with the higher-order measurement model (3.15) at the cost of this strengthening of the independence assumption. A preliminary step before the closed-form identification of $g(X^*)$ involves identification of the polynomial coefficients $\gamma_0, \cdots, \gamma_P$ and the moments of ϵ_2 up to the P-th order. This preliminary step is presented in Section 3.2.3. After the preliminary step, we then proceed with closed-form identification of the nonparametric regression function g in Section 3.2.3.

A Preliminary Step: Identification of γ_p and $E[\epsilon_2^p]$

As is the case for the simple affine model of endogenous measurement presented in Section 3.2.3 (see (3.13) and Lemma 3.2.1), identification of the parameters γ_p and $\sigma_2^p := E[\epsilon_2^p]$ for the model (3.15) also follows from an appropriate set of moment restrictions. To form such restrictions, one can propose several alternative statistical and mean independence assumptions, and there is not the unique set of identifying restrictions to this goal. One might therefore want to come up with the most convenient set of restriction tailored to specific empirical applications. As a general prescription, we can form restrictions of the form

$$\begin{aligned}
cov(YX_2^q, X_1) &= E\left[Y(X^*+\epsilon_2)^q\left(\sum_{p=1}^{P}\gamma_p X^{*p}+\epsilon_1\right)\right] - E\left[Y(X^*+\epsilon_2)^q\right]E\left[\sum_{p=1}^{P}\gamma_p X^{*p}+\epsilon_1\right] \\
&= \sum_{p=0}^{P}\sum_{q'=0}^{q}\gamma_p\sigma_2^{q-q'}\binom{q}{q'}\left(E[YX^{*(p+q')}] - E[YX^{*q'}]E[X^{*p}]\right) \\
cov(YX_2^r, X_2^s) &= E[Y(X^*+\epsilon_2)^{r+s}] - E[Y(X^*+\epsilon_2)^r]E[(X^*+\epsilon_2)^s] \\
&= \sum_{r'=0}^{r+s}\sigma_2^{r+s-r'}\binom{r+s}{r'}E[YX^{*r'}] - \sum_{r'=0}^{r}\sum_{s'=0}^{s}\sigma_2^{r+s-r'-s'}\binom{r}{r'}\binom{s}{s'}E[YX^{*r'}]E[X^{*s'}]
\end{aligned}$$

for various $q = 0, 1, \cdots, Q-P$, $r = 0, 1, \cdots$ and $s = 1, \cdots$ such that $r + s \leqslant Q$ for some $Q \in \mathbb{N}$. The right-hand sides of the above two equations involve the unknowns, $(\gamma_0, \cdots, \gamma_P)$, $(\sigma_2^2, \cdots, \sigma_2^Q)$, $(E[X^*], \cdots, E[X^{*Q}])$, and $(E[YX^*], \cdots, E[YX^{*Q}])$, under Assumption 3.2.5. As such, we obtain $(Q - P + 1) + \frac{Q(Q+1)}{2}$ restrictions for $3Q + P$ unknown parameters, $(\gamma_0, \cdots, \gamma_P)$, $(\sigma_2^2, \cdots, \sigma_2^Q)$, $(E[X^*], \cdots, E[X^{*Q}])$, and $(E[YX^*], \cdots, E[YX^{*Q}])$. Clearly for any given order P of polynomial, as we increase the number Q, we have sufficiently more

number of restrictions than the unknowns to recover the polynomial coefficients $\gamma_0, \cdots, \gamma_P$ and the moments $\sigma_2^2, \cdots, \sigma_2^P$ which we need.

A drawback to the above general prescription is that these moment restrictions may not necessarily lead to a closed-form solution to these parameters. One can make alternative statistical and mean independence assumptions for the goal of obtaining closed-form identification of the polynomial coefficients $\gamma_0, \cdots, \gamma_P$ and the moments $\sigma_2^2, \cdots, \sigma_2^P$. Specifically, we may show a closed-from solution to the quadratic case, where the endogenous measurement X_1 is modeled with $P = 2$ by

$$X_1 = \gamma_1 X^* + \gamma_2 X^{*2} + \epsilon_1 \ : \qquad E[\epsilon_1] = \gamma_0 \tag{3.16}$$

We impose a homoscedasticity assumption

$$E[U^2 \mid X^*, \epsilon_1, \epsilon_2] = E[U^2] \tag{3.17}$$

and an empirically testable rank condition

$$cov(Y, X_2) \cdot cov(Y^2, X_2^2) \neq cov(Y, X_2^2) \cdot cov(Y^2, X_2). \tag{3.18}$$

We may then show that the coefficients γ_1 and γ_2 of the model (3.16) are identified with closed-form solutions as follows.

$$\gamma_1 = \frac{cov(Y, X_1) \cdot cov(Y^2, X_2^2) - cov(Y, X_2^2) \cdot cov(Y^2, X_1)}{cov(Y, X_2) \cdot cov(Y^2, X_2^2) - cov(Y, X_2^2) \cdot cov(Y^2, X_2)} \tag{3.19}$$

$$\gamma_2 = \frac{cov(Y, X_2) \cdot cov(Y^2, X_1) - cov(Y, X_1) \cdot cov(Y^2, X_2)}{cov(Y, X_2) \cdot cov(Y^2, X_2^2) - cov(Y, X_2^2) \cdot cov(Y^2, X_2)}, \tag{3.20}$$

Furthermore, Assumption 3.2.5 also allows us to identify γ_0 and σ_2^2 with closed-form solutions from the system of linear equations as follows:

$$\begin{bmatrix} E[YX_1] \\ E[X_1X_2] \\ E[YX_1X_2] \end{bmatrix} = \begin{bmatrix} E[Y] & -\gamma_2 E[Y] & 0 \\ E[X_2] & -\gamma_1 - 3\gamma_2 E[X_2] & -\gamma_2 \\ E[YX_2] & -\gamma_1 E[Y] - 3\gamma_2 E[YX_2] & -\gamma_2 E[Y] \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \sigma_2^2 \\ \sigma_2^3 \end{bmatrix}.$$

If the above 3 by 3 matrix is nonsingular, the linear system yields a unique solution to $(\gamma_0, \sigma_2^2, \sigma_2^3)$. In particular, it yields the following closed-form formula for σ_2^2:

$$\sigma_2^2 = \frac{1}{2\gamma_2} \left(\frac{cov(Y, X_1X_2)}{cov(Y, X_2)} - \frac{E[YX_1]}{E[Y]} \right).$$

Identification of Nonparametric Regression Function

Having identified the polynomial coefficients $(\gamma_1, \cdots, \gamma_P)$ and the moments $(\sigma_2^2, \cdots, \sigma_2^P)$ for the model (3.15) following the methods outlined in Section 3.2.3, we proceed with closed-form identification of the nonparametric regression function g evaluated at various points x^* in the interior of the support of X^*. To this end, we assume the following rank condition,

which is effectively an empirically testable assumption as $(\sigma_2^2, \cdots, \sigma_2^P)$ are identified from observed data $F_{YX_1X_2}$.

Assumption 3.2.6 (Empirically Testable Rank Condition) *The following matrix is nonsingular.*

$$\begin{bmatrix} 1 & \binom{P}{P-1}\sigma_2^1 & \cdots & \binom{P}{2}\sigma_2^{P-2} & \binom{P}{1}\sigma_2^{P-1} \\ & 1 & \cdots & \binom{P-1}{2}\sigma_2^{P-3} & \binom{P-1}{1}\sigma_2^{P-2} \\ & & \ddots & \vdots & \vdots \\ & & & 1 & \binom{2}{1}\sigma_2^1 \\ & & & & 1 \end{bmatrix}_{P\times P}$$

Besides its empirical testability, this rank condition is automatically satisfied for the linear case ($P = 1$) and the quadratic case ($P = 2$) due to the normalization $E[\epsilon_2] = 0$ in (3.15).[6] For convenience of writing, we let Z^* denote the random variable $\sum_{p=1}^P \gamma_p X^{*p}$. The role of Assumption 3.2.6 is to identify the distribution of this generated latent variable Z^* in the following manner. Under Assumption 3.2.6, we can write the following vector on the left-hand side in terms of the expression on the right-hand side that consists of observed data.

$$\begin{bmatrix} \mu(t, P; \sigma_2^1, \cdots, \sigma_2^P; F_{X_1X_2}) & \cdots & \mu(t, 1; \sigma_2^1, \cdots, \sigma_2^P; F_{X_1X_2}) \end{bmatrix}' := \\ \begin{bmatrix} 1 & \binom{P}{P-1}\sigma_2^1 & \cdots & \binom{P}{2}\sigma_2^{P-2} & \binom{P}{1}\sigma_2^{P-1} \\ & 1 & \cdots & \binom{P-1}{2}\sigma_2^{P-3} & \binom{P-1}{1}\sigma_2^{P-2} \\ & & \ddots & \vdots & \vdots \\ & & & 1 & \binom{2}{1}\sigma_2^1 \\ & & & & 1 \end{bmatrix}^{-1} \begin{bmatrix} E[(X_2^P - \sigma_2^P)e^{itX_1}] \\ E[(X_2^{P-1} - \sigma_2^{P-1})e^{itX_1}] \\ \vdots \\ E[(X_2^2 - \sigma_2^2)e^{itX_1}] \\ E[(X_2 - \sigma_2^1)e^{itX_1}] \end{bmatrix} \tag{3.21}$$

It is shown in the theorem below that this vector is sufficient to pin down the distribution of the generated latent variable $Z^* = \sum_{p=1}^P \gamma_p X^{*p}$, and hence its distribution (equivalently, its characteristic function) can be identified from observed data.

To make use of this auxiliary result to identify the nonparametric regression function g of interest, we next propose the following regularity conditions.

Assumption 3.2.7 (Regularity) *(i) ϕ_{X_1} and ϕ_{X_2} do not vanish on the real line. (ii) f_{X^*} and $\mathcal{F}f_{X^*}$ are continuous and absolutely integrable. (iii) f_{Z^*} and $\mathcal{F}f_{Z^*}$ are continuous and absolutely integrable. (iv) $f_{X^*} \cdot g$ and $\mathcal{F}(f_{X^*} \cdot g)$ are continuous and absolutely integrable.*

This assumption plays a similar role to Assumption 3.2.4. In parts of our identifying formula, the characteristic functions appear as denominators, and hence part (i) of this assumption rules out zero denominator. This property of non-vanishing characteristic functions is shared by many of the common distribution families, e.g., the normal, chi-squared, Cauchy, gamma, and exponential distributions. Parts (ii) and (iii) ensure that we can apply

[6] However, when the order of polynomial is $P = 3$ or above, this rank condition can be shown to be unsatisfied, e.g., one can check that $\sigma_2^2 = \frac{1}{3}$ when $P = 3$ fails the assumption.

the Fourier transform and inversion to those functions. The model allows for nonlinear and endogenous errors in the sense of $E[X_1 \mid X^*] = \sum_{p=0}^{P} \gamma_p X^{*p}$. However, we rule out the case where the report X_1 is decreasing while the truth X^* is increasing. Specifically, we assume the following monotonicity restriction.

Assumption 3.2.8 (Monotonicity) *$\sum_{p=0}^{P} \gamma_p x^p$ is non-decreasing in x on the support of X^*.*

This monotonicity assumption is used for the purpose of applying the density transformation formula to derive the density function for the transformed random variable. Polynomial functions do not generally exhibit monotonicity on the entire real line. Note that Assumption 3.2.8 only requires the monotonicity to hold on the support of X^*, and hence is not restrictive when the support of X^* is a proper subset of $\mathbb{R}$. For example, many economic variables X^* are innately positive, i.e., $\text{supp}(X^*) \subseteq \mathbb{R}_+$, and the quadratic function $E[X_1 \mid X^*] = \gamma_2 X^{*2}$, for example, necessarily satisfies Assumption 3.2.8 for such variables.

With this set of assumptions, we can still identify the nonparametric function g with a closed-form formula, even if the measurement X_1 is systematically biased with endogeneity and such a high order of nonlinearity. The following theorem states the exact result.

Theorem 3.2.2 *Suppose that Assumptions 3.2.5, 3.2.6, 3.2.7 and 3.2.8 hold for the model (3.15). The nonparametric function g evaluated at x^* in the interior of the support of X^* is identified with the closed-form solution:*

$$g(x^*) = \frac{\int\int\int e^{-itx^*+itx-it'\left(\sum_{p=1}^{P}\gamma_p x^p\right)} \left|\sum_{p=1}^{P} p\gamma_p x^{p-1}\right| \frac{E[Ye^{itX_2}]}{E[e^{itX_2}]} \phi_{Z^*}(t')dt'dxdt}{2\pi \int e^{-ith\left(\sum_{p=1}^{P}\gamma_p x^{*p}\right)} \left|\sum_{p=1}^{P} p\gamma_p x^{*(p-1)}\right| \phi_{Z^*}(t)dt},$$

where ϕ_{Z^*} is identified with the closed-form solution

$$\phi_{Z^*}(t) = \exp\left\{\int_0^t \frac{\sum_{p=1}^{P}\gamma_p \mu(t_1, p\ ; \sigma_2^1, \cdots, \sigma_2^P; F_{X_1X_2})}{E[e^{it_1X_1}]} dt_1\right\}$$

and $\mu(t, p\ ; \sigma_2^1, \cdots, \sigma_2^P; F_{X_1X_2})$ for all $p = 1, \cdots, P$ are given by the closed-form solution (3.21).

Note that this general version of the closed-form identifying formula, involving the triple integral instead of a single integral due to the nonlinear transformation, is qualitatively quite different from the traditional formulas including the one in Theorem 3.2.1 as well as that of Schennach (2004b). Theorem 3.2.1 may appear to be a special case of this theorem, as the former focuses on affine models and the latter extends to higher order polynomials. Strictly speaking, it is not a special case, because Theorem 3.2.1 requires slightly weaker independence assumptions than Theorem 3.2.2. As such, we stated Theorem 3.2.1 separately in the previous section for the practical importance of parsimonious affine models.

Closed-Form Estimator

Given the closed-form identifying formulas of Theorems 3.2.1 and 3.2.2, one can easily construct a direct sample-counterpart estimator by replacing the population moments by the sample moments for the characteristic functions. As this basic idea is the same across all the cases, we focus on the simplest model (3.12) for simplicity in this section. If γ_1 is known, then the sample-counterpart estimator $\widehat{g(x^*)}$ of the closed-form identifying formula (3.14) is given by

$$\widehat{g(x^*)} = \frac{\int_{-\infty}^{+\infty} e^{-itx^*} \exp\left(i \int_0^t \frac{\sum_{j=1}^n X_{2,j} e^{it_1 X_{1,j}/\gamma_1}}{\sum_{j=1}^n e^{it_1 X_{1,j}/\gamma_1}} dt_1 \right) \frac{\sum_{j=1}^n Y_j e^{itX_{1,j}/\gamma_1}}{\sum_{j=1}^n e^{itX_{1,j}/\gamma_1}} \phi_K(th) dt}{\int_{-\infty}^{+\infty} e^{-itx^*} \exp\left(i \int_0^t \frac{\sum_{j=1}^n X_{2,j} e^{it_1 X_{1,j}/\gamma_1}}{\sum_{j=1}^n e^{it_1 X_{1,j}/\gamma_1}} dt_1 \right) \phi_K(th) dt} \tag{3.22}$$

where ϕ_K denotes the Fourier transform of a kernel function K which we use together with the tuning parameter h for the purpose of regularization.

On the other hand, if γ_1 is not known, we replace γ_1 by its estimate and the estimator thus takes the form

$$\widehat{g(x^*)} = \frac{\int_{-\infty}^{+\infty} e^{-itx^*} \exp\left(i \int_0^t \frac{\sum_{j=1}^n X_{2,j} e^{it_1 X_{1,j}/\hat{\gamma}_1}}{\sum_{j=1}^n e^{it_1 X_{1,j}/\hat{\gamma}_1}} dt_1 \right) \frac{\sum_{j=1}^n Y_j e^{itX_{1,j}/\hat{\gamma}_1}}{\sum_{j=1}^n e^{itX_{1,j}/\hat{\gamma}_1}} \phi_K(th) dt}{\int_{-\infty}^{+\infty} e^{-itx^*} \exp\left(i \int_0^t \frac{\sum_{j=1}^n X_{2,j} e^{it_1 X_{1,j}/\hat{\gamma}_1}}{\sum_{j=1}^n e^{it_1 X_{1,j}/\hat{\gamma}_1}} dt_1 \right) \phi_K(th) dt} \tag{3.23}$$

where $\hat{\gamma}_1$ is computed by the following sample-counterpart of (3.13).

$$\hat{\gamma}_1 = \frac{\frac{1}{n}\sum_{j=1}^n Y_j X_{1,j} - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)\left(\frac{1}{n}\sum_{j=1}^n X_{1,j}\right)}{\frac{1}{n}\sum_{j=1}^n Y_j X_{2,j} - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)\left(\frac{1}{n}\sum_{j=1}^n X_{2,j}\right)}.$$

It turns out that the substitution of the estimate $\hat{\gamma}_1$ for the true value of γ_1 does not affect the asymptotic property of $\widehat{g(x^*)}$. We assume the following basic regularity conditions to derive the consistency of $\widehat{g(x^*)}$ in both (3.22) and (3.23).

Assumption 3.2.9 (Basic Assumptions for Consistency of the Closed-Form Estimator)

(i) $\{X^*, \epsilon_1, \epsilon_2, U\}$ *is independently and identically distributed.*
(ii) ϕ_K *is symmetric, satisfies* $\phi_K(0) = 1$*, and has integrable second derivatives.*
(iii) $E|X_1|^{2+\delta} < \infty$*,* $E|X_2|^{2+\delta} < \infty$*, and* $E|Y|^{2+\delta} < \infty$ *for some* $\delta > 0$*.*

In case of using the version (3.23) of the closed-form estimator instead of (3.22), we assume the following bounded fourth moment restriction in addition to part (iii) of Assumption 3.2.9.

Assumption 3.2.9. *(iii)′* $E|X_1|^4 < \infty$*,* $E|X_2|^4 < \infty$*, and* $E|Y|^4 < \infty$*.*

The asymptotic rate of convergence of the closed-form estimators (3.22) and (3.23) depend

on the Hölder exponents of the nonparametric density f_{X^*} and the nonparametric regression g. We therefore introduce the following assumption with index numbers that determine the asymptotic orders.

Assumption 3.2.10 (Determinants of the Asymptotic Orders of Biases) *(i) f_{X^*} is twice continuously differentiable at x^*, and the k_1-th derivative of f_{X^*} is k_2-Hölder continuous with Hölder constant bounded by k_0, i.e.,*

$$\left|f_{X^*}^{(k_1)}(x) - f_{X^*}^{(k_1)}(x+\delta)\right| \leqslant k_0 \left|\delta\right|^{k_2} \qquad \text{for all } x \text{ and } \delta.$$

(ii) g is twice continuously differentiable at x^*, and the l_1-th derivative of g is l_2-Hölder continuous with Hölder constant bounded by l_0, i.e.,

$$\left|g^{(l_1)}(x) - g^{(l_1)}(x+\delta)\right| \leqslant l_0 \left|\delta\right|^{l_2} \qquad \text{for all } x \text{ and } \delta.$$

Let $k = k_1 + k_2$ and $l = l_1 + l_2$ be the largest numbers satisfying the above properties.

Since optimal choices of the bandwidth parameter h depend on the shape of the underlying characteristic function, we first state the following auxiliary result of convergence rate under free choice of h.

Lemma 3.2.2 (Mean Square Error of the Closed-Form Estimator) *Suppose that Assumptions 3.2.2, 3.2.3 and 3.2.4 hold for the model (3.12). If Assumptions 3.2.9 and 3.2.10 are satisfied and x^* is in the interior of the support of X^*, then, with any choice of h such that $h \to 0$ and $nh^4 \left|\phi_{X_1}(1/h)\right|^4 \to \infty$ as $n \to \infty$, the mean square error of the closed-form estimator $\widehat{g(x^*)}$ given in (3.22) has the asymptotic order:*

$$\mathcal{O}(h^{2\min\{k,l\}}) + \mathcal{O}\left(\frac{1}{nh^4 \left|\phi_{X_1}(1/h)\right|^4}\right), \tag{3.24}$$

where the first and second terms correspond to the asymptotic orders of the squared bias and the variance, respectively. The same conclusion holds for the closed-form estimator $\widehat{g(x^*)}$ given in (3.23), provided that Assumptions 3.2.1, and 3.2.9 (iii)′ additionally hold.

This lemma implies that the MSE-optimizing choice of h obviously depends on the tail behavior of the characteristic function ϕ_{X_1}, which in turn depends on the characteristic functions ϕ_{X^*} and ϕ_{ϵ_1}. Therefore, we branch into the following two cases: (a) at least one of X^* and ϵ_1 has a super-smooth distribution; and (b) both X^* and ϵ_1 have ordinary-smooth distributions. These two cases are precisely stated in the following two separate assumptions.

Assumption 3.2.11 (Super-Smooth Distributions) *Assume that*
(i) the distribution of X^* is super-smooth of order $\beta_1 > 0$, i.e., there exist $\kappa_1 > 0$ such that

$$\left|\phi_{X^*}(t)\right| = \mathcal{O}\left(e^{-\left|t\right|^{\beta_1}/\kappa_1}\right) \qquad \text{as } t \to \pm\infty,$$

OR
(ii) the distribution of ϵ_1 is super-smooth of order $\beta_2 > 0$, i.e., there exist $\kappa_2 > 0$ such that

$$|\phi_{\epsilon_1}(t)| = \mathcal{O}\left(e^{-|t|^{\beta_2}/\kappa_2}\right) \quad \text{as } t \to \pm\infty,$$

OR both (i) and (ii) hold. For convenience of notation, we let $\beta_1 = 0$ (respectively, $\beta_2 = 0$) if the distribution of X^* (respectively, ϵ_1) is not super-smooth.

Assumption 3.2.12 (Ordinary-Smooth Distributions) *Assume that*
(i) the distribution of X^* is ordinary-smooth of order β_1, i.e.,

$$|\phi_{X^*}(t)| = \mathcal{O}\left(|t|^{-\beta_1}\right) \quad \text{as } t \to \pm\infty,$$

AND
(ii) the distribution of ϵ_1 is ordinary-smooth of order β_2, i.e.,

$$|\phi_{\epsilon_1}(t)| = \mathcal{O}\left(|t|^{-\beta_2}\right) \quad \text{as } t \to \pm\infty.$$

These two smoothness definitions characterized by the tail behavior of the characteristic functions measure the smoothness of the density function. Examples of super-smooth distributions include the normal, Cauchy, and mixed normal distributions. Examples of ordinary-smooth distributions include the gamma, exponential, and uniform distributions. If at least one of X^* and ϵ_1 has a super-smooth distribution in the sense of Assumption 3.2.11, then the closed-form estimators follow $\log n$ rates of convergence as follows.

Theorem 3.2.3 (Consistency under Super-Smooth Distribution(s)) *Suppose that Assumptions 3.2.2, 3.2.3 and 3.2.4 hold for the model (3.12). If Assumptions 3.2.9, 3.2.10, and 3.2.11 are satisfied and x^* is in the interior of the support of X^*, then the closed-form estimator $\widehat{g(x^*)}$ given in (3.22) is consistent with the convergence rate*

$$\left(E\left[\widehat{g(x^*)} - g(x^*)\right]^2\right)^{1/2} = \mathcal{O}\left((\log n)^{\frac{-\min\{k,l\}}{\max\{\beta_1,\beta_2\}}}\right)$$

under the choice of the tuning parameter $h \propto (\log n)^{-1/\max\{\beta_1,\beta_2\}}$. The same conclusion holds for the closed-form estimator $\widehat{g(x^*)}$ given in (3.23), provided that Assumptions 3.2.1, and 3.2.9 (iii)′ additionally hold.

On the other hand, if both X^* and ϵ_1 have ordinary-smooth distributions in the sense of Assumption 3.2.12, then the closed-form estimator follow polynomial rates of convergence as follows.

Theorem 3.2.4 (Consistency under Ordinary-Smooth Distributions) *Suppose that Assumptions 3.2.2, 3.2.3 and 3.2.4 hold for the model (3.12). If Assumptions 3.2.9, 3.2.10, and 3.2.12 are satisfied and x^* is in the interior of the support of X^*, then the closed-form*

estimator $\widehat{g(x^*)}$ *given in (3.22) is consistent with the convergence rate*

$$\left(E\left[\widehat{g(x^*)} - g(x^*)\right]^2\right)^{1/2} = \mathcal{O}\left(n^{\frac{-\min\{k,l\}}{2(\min\{k,l\}+2(\beta_1+\beta_2+1))}}\right)$$

under the choice of the tuning parameter $h \propto n^{-1/2(\min\{k,l\}+2(\beta_1+\beta_2+1))}$. *The same conclusion holds for the closed-form estimator* $\widehat{g(x^*)}$ *given in (3.23), provided that Assumptions 3.2.1, and 3.2.9 (iii)′ additionally hold.*

While the contexts and the setups are different and a direct comparison cannot be made, the two cases covered in our Theorems 3.2.3 and 3.2.4 can be connected to Cases 2 and 4 of Theorem 2 in Schennach (2004b), respectively.[7] The rates of convergence achieved by the estimator in both cases fall short of the convergence rate of the traditional nonparametric regression estimators that assume observation of X^*.

[7]Specifically, the auxiliary parameters β_v, γ_b and γ_v used in Schennach (2004b) can be reconciled with our regularity parameters through the relations $\beta_v = \max\{\beta_1, \beta_2\}$, $\gamma_b = -\min\{k, l\}$ and $\gamma_v = 2(\beta_1 + \beta_2 + 1)$.

4

Applications in Empirical Industrial Organization

A major breakthrough in the measurement error literature is the nonparametric identification of the 2.1-measurement model in section 2.4, which allows a very flexible relationship between observables and unobservables. The generality of these results enables researchers to tackle many important problems involving latent variables, such as belief, productivity, unobserved heterogeneity, and fixed effects, in the field of empirical industrial organization and labor economics.

4.1 Unobserved Heterogeneity in Auctions

Unobserved heterogeneity has been a concern in the estimation of auction models for a long time. Li et al. (2000) and Krasnokutskaya (2011) use the identification result of 2-measurement model in equation (2.30) to estimate auction models with separable unobserved heterogeneity. In a first-price auction indexed by t for $t = 1, 2, \ldots, T$ without a reserve price, there are N symmetric risk-neutral bidders. For $i = 1, 2, \ldots, N$, each bidder i's cost is assumed to be decomposed into two independent factors as $s_t^* \times x_i$, where x_i is her private value and s_t^* is an auction-specific state or unobserved heterogeneity. With this decomposition of the cost, it can be shown that equilibrium bidding strategies b_{it} can also be decomposed as follows

$$b_{it} = s_t^* a_i, \tag{4.1}$$

where $a_i = a_i(x_i)$ represents equilibrium bidding strategies in the auction with $s_t^* = 1$. This falls into the 2-measurement model given that

$$b_{1t} \perp b_{2t} \mid s_t^*. \tag{4.2}$$

With such separable unobserved heterogeneity, one can consider the joint distribution of two bids as follows:

$$\begin{aligned} \ln b_{1t} &= \ln s_t^* + \ln a_1 \\ \ln b_{2t} &= \ln s_t^* + \ln a_2, \end{aligned} \tag{4.3}$$

where Kotlarski's identity is applicable for nonparametric identification of the distributions of $\ln s_t^*$ and $\ln a_i$. Further estimation of the value distribution from the distribution of a_i (x_i) can be found in Guerre et al. (2000) .

Hu et al. (2013a) consider auction models with non-separable unobserved heterogeneity. They assume that the private values x_i are independent conditional on an auction-specific state or unobserved heterogeneity s_t^*. Based on the conditional independence of the values, the conditional independence of the bids holds, i.e.,

$$b_{1t} \perp b_{2t} \perp b_{3t} \mid s_t^*. \tag{4.4}$$

This falls into a 3-measurement model, where the three measurements, i.e., bids, are independent conditional on the unobserved heterogeneity. Nonparametric identification of the model then follows.

A specific example of unobserved heterogeneity may be bidder's heterogeneous beliefs across Auctions. An (2017) considers first-price auctions, in which bidders' beliefs about their opponents' bidding behavior are not in equilibrium but follow a level-k thinking as in Stahl and Wilson (1994). Bidders are assumed to have different levels of sophistication with a hierarchical structure, i.e., heterogenous (possibly incorrect) beliefs about others' behavior based on a nonstrategic type as follows.

Type	Belief about other bidders' behavior
1	all other bidders are type-$L0$ (bid naïvely)
2	all other bidders are type-1
⋮	⋮
k	all other bidders are type-$(k-1)$

By observing a bidder's behavior in three different auctions, An (2017) uses a 3-measurement model to show the model with latent belief levels (or types) can be identified and estimated. The key assumptions include that bidder's belief level doesn't change across auctions and that three bids are independent conditional on the belief level. Such an empirical model helps explain overbidding and non-equilibrium behavior. More detailed description can be found in Yonghong An's presentation slides ↗.

4.2 Auctions with an Unknown Number of bidders

Since the earliest papers in the structural empirical auction literature, researchers have had to grapple with a lack of information on N^*, the number of potential bidders in the auction,

which is an indicator of market competitiveness. The number of potential bidders may be different from the observed number of bidders A due to binding reserve prices, participation costs, or misreporting errors. For example, when reserve prices are binding, the number of potential bidders N^* would be observed by bidders and affect their bidding behavior. However, the observed number of bidders A, which is the number of participants whose bids exceed the reserve price, would be less than or equal to N^*.

In first-price sealed-bid auctions under the symmetric independent private values (IPV) paradigm, each of N^* potential bidders draws a private valuation from the distribution $F_{N^*}(x)$ with support $[\underline{x}, \overline{x}]$. The bidders observe N^*, which is latent to researchers. The reserve price r is assumed to be known and fixed across all auctions with $r > \underline{x}$. For each bidder i with valuation x_i, the equilibrium bidding function $b(x_i, N^*)$ can be shown as follows:

$$b(x_i; N^*) = \begin{cases} x_i - \frac{\int_r^{x_i} F_{N^*}(s)^{N^*-1} ds}{F_{N^*}(x_i)^{N^*-1}} & \text{for} \quad x_i \geq r \\ 0 & \text{for} \quad x_i < r. \end{cases} \tag{4.5}$$

The observed number of bidders is $A_t = \sum_{i=1}^{N_t^*} \mathbf{1}(x_{it} \geq r)$. In a random sample, we observe $\{A_t, b_{1t}, b_{2t}, \ldots, b_{A_t t}\}$ for each auction $t = 1, 2, \ldots, T$. We consider

$$\begin{aligned}
& f(b_{1t}, b_{2t}, A_t, x_{1t}, x_{2t} | x_{1t} \geq r, x_{2t} \geq r, N_t^*) \\
= \; & f(b_{1t} | b_{2t}, A_t, x_{1t}, x_{2t}, x_{1t} \geq r, x_{2t} \geq r, N_t^*) f(b_{2t} | A_t, x_{1t}, x_{2t}, x_1 \geq r, x_{2t} \geq r, N_t^*) \\
& \times f(A_t | x_{1t}, x_{2t}, x_{1t} \geq r, x_{2t} \geq r, N_t^*) f(x_{1t}, x_{2t} | x_{1t} \geq r, x_{2t} \geq r, N_t^*) \\
= \; & f(b_{1t} | x_{1t}, x_{1t} \geq r, N_t^*) f(b_{2t} | x_{2t}, x_{2t} \geq r, N_t^*) \\
& \times f\left(\sum_{i=3}^{N^*} \mathbf{1}(x_{it} \geq r) + \mathbf{1}(x_{1t} \geq r) + \mathbf{1}(x_{2t} \geq r) | x_{1t}, x_{2t}, x_{1t} \geq r, x_{2t} \geq r, N_t^* \right) \\
& \times f(x_{1t}, x_{2t} | x_{1t} \geq r, x_{2t} \geq r, N_t^*) \\
= \; & f(b_{1t} | x_{1t}, x_{1t} \geq r, N_t^*) f(b_{2t} | x_{2t}, x_{2t} \geq r, N_t^*) \\
& \times f\left(\sum_{i=3}^{N^*} \mathbf{1}(x_{it} \geq r) + 2 | N_t^* \right) f(x_{1t} | x_{2t}, x_{1t} \geq r, x_{2t} \geq r, N_t^*) f(x_{2t} | x_{1t} \geq r, x_{2t} \geq r, N_t^*) \\
= \; & f(b_{1t} | x_{1t}, x_{1t} \geq r, N_t^*) f(b_{2t} | x_{2t}, x_{2t} \geq r, N_t^*) f(A_t | A_t \geq 2, N_t^*) \\
& \times f(x_{1t} | x_{1t} \geq r, N_t^*) f(x_{2t} | x_{2t} > r, N_t^*)
\end{aligned}$$

Note that the event $\{x_{it} \geq r\}$ is the same as $\{b_{it} \geq r\}$. We have

$$\begin{aligned}
& f(b_{1t}, b_{2t}, A_t, x_{1t}, x_{2t} | b_{1t} \geq r, b_{2t} \geq r, N_t^*) \\
= \; & f(b_{1t} | x_{1t}, b_{1t} \geq r, N_t^*) f(b_{2t} | x_{2t}, b_{2t} \geq r, N_t^*) f(A_t | A_t \geq 2, N_t^*) \\
& \times f(x_{1t} | b_{1t} \geq r, N_t^*) f(x_{2t} | b_{2t} > r, N_t^*)
\end{aligned}$$

Integrating out x_{1t}, x_{2t} leads to

$$\begin{aligned} & f\left(b_{1t}, b_{2t}, A_t | b_1 \geq r, b_2 \geq r, N^*\right) \\ = & \int f\left(b_{1t} | x_{1t}, b_{1t} \geq r, N_t^*\right) f\left(x_{1t} | b_{1t} \geq r, N_t^*\right) dx_{1t} \\ & \times \int f\left(b_{2t} | x_{2t}, b_{2t} \geq r, N_t^*\right) f\left(x_{2t} | b_{2t} > r, N_t^*\right) dx_{2t} \\ & \times f\left(A_t | A_t \geq 2, N_t^*\right) \\ = & f\left(b_{1t} | b_{1t} \geq r, N^*\right) f\left(b_{2t} | b_{2t} \geq r, N^*\right) f\left(A_t | A_t \geq 2, N^*\right). \end{aligned}$$

Therefore, we have

$$\begin{aligned} & f\left(A_t, b_{1t}, b_{2t} | b_{1t} > r, b_{2t} > r\right) \\ = & \sum_{N^*} f\left(A_t | A_t \geq 2, N^*\right) f\left(b_{1t} | b_{1t} > r, N^*\right) f\left(b_{2t} | b_{2t} > r, N^*\right) f\left(N^* | b_{1t} > r, b_{2t} > r\right). \end{aligned} \tag{4.6}$$

That means that the two bids and the observed number of bidders are independent conditional on the number of potential bidders, which forms a 3-measurement model. In addition, the fact that $A_t \leq N_t^*$ provides an ordering of the eigenvectors corresponding to $f_{A_t|N_t^*}$. As shown in An et al. (2010) , the bid distribution, and therefore, the value distribution, can be non-parametrically identified. Furthermore, such identification is constructive and directly leads to an estimator.

An et al. (2010) consider nonparametric identification and estimation of first-price auction models when N^* is observed by bidders, but not by the researcher. Using recent results from the literature on misclassified regressors, we show how the equilibrium distribution of bids, given the unobserved N^*, can be identified and estimated. In the case of first-price auctions, these bid distributions estimated using our procedure can be used as inputs into established nonparametric procedures (Guerre et al. (2000), Li et al. (2002)) to obtain estimates of bidders' valuations.

Accommodating the possibility that the researcher does not know N^* is important for drawing valid policy implications from auction model estimates. Because N^* is the level of competition in an auction, not knowing N^*, or using a mismeasured value for N^*, can lead to wrong implications about the degree of competitiveness in the auction, and also the extent of bidders' markups and profit margins. Indeed, a naïve approach where the number of observed bids is used as a proxy for N^* will tend to overstate competition, because the unknown N^* is always (weakly) larger than the number of observed bids. This bias will be shown in the empirical illustration below.

Not knowing the potential number of bidders N^* has been an issue since the earliest papers in the structural empirical auction literature. In the parametric estimation of auction models, the functional relationship between the bids b and number of potential bidders N^* is explicitly parameterized, so that not knowing N^* need not be a problem. For instance, Laffont et al. (1995) used a goodness-of-fit statistic to select the most plausible value of N^* for French eggplant auctions. Paarsch (1997) treated N^* essentially as a random effect and integrates it out over the assumed distribution in his analysis of timber auctions.

In a nonparametric approach to auctions, however, the relationship between the bids b and N^* must be inferred directly from the data, and not knowing N^* (or observing N^* with error) raises difficulties. Within the independent private-values (IPV) framework, and under the additional assumption that the unknown N^* is fixed across all auctions (or fixed across a known subset of the auctions), Guerre et al. (2000) showed how to identify N^* and the equilibrium bid distribution in the range of bids exceeding the reserve price. Hendricks et al. (2003) allowed N^* to vary across auctions, and assume that $N^* = L$, where L is a measure of the number of potential bidders which they construct.

The main contribution of An et al. (2010) is to present a solution for the nonparametric identification and estimation of first-price auction models in which the number of bidders N^* is observed by bidders, but unknown to the researcher. We develop a nonparametric procedure for recovering the distribution of bids conditional on unknown N^* which requires neither N^* to be fixed across auctions, nor for an (assumed) perfect measure of N^* to be available. Our procedure applies results from the recent econometric literature on models with misclassification error, such as e.g. Mahajan (2006), Hu (2008).

As a specific case, our method is, as far as we aware, the first to solve the identification problem for IPV first-price auctions with reserve prices when the unobserved number of potential bidders N^* is a random variable. Previously, Guerre et al. (2000) also considered identification for first-price IPV auctions with reserve prices. However, they assumed that the observed number of potential bidders N^* is fixed across auctions, so that it could be estimated as a parameter.

For first-price auctions, allowing the unknown N^* to vary randomly across auctions is not innocuous. Because N^* is observed by the bidders, it affects their equilibrium bidding strategies. Hence, when N^* is not known by the researcher, and varies across auctions, the observed bids are drawn from a mixture distribution, where the "mixing densities" $g(b|N^*)$ and the "mixing weights" $\Pr(A|N^*)$ are both unknown. This motivates the application of econometric methods developed for models with a misclassified regressor, where (likewise) the observed outcomes are drawn from a mixture distribution.

Most closely related to our work is a paper by Song (2004). She solved the problem of the nonparametric estimation of ascending auction models in the IPV framework, when the number of potential bidders N^* is unknown by the researcher (and varies in the sample). She showed that the distribution of valuations can be recovered from observation of any two valuations of which rankings from the top is known.[1] However, her approach cannot be applied to first-price auctions, which are the focus of this paper. The reason for this is that, in IPV first-price auctions (but not in ascending- or second-price auctions), even if the distribution of bidders' valuations do not vary across the unknown N^*, the equilibrium distribution of bids still vary across N^*. Hence, because the researcher does not know N^*, the observed bids are drawn from a mixture distribution, and estimating the model requires deconvolution methods which have been developed in the econometric literature on measurement error.[2]

[1] Adams (2007) also considers estimation of ascending auctions when the distribution of potential bidders is unknown.

[2] Song (2006) showed that the top two bids are also enough to identify first-price auctions where the

In a different context, Li et al. (2000) applied deconvolution results from the (continuous) measurement error literature to identify and estimate conditionally independent auction models in which bidders' valuations have common and private (idiosyncratic) components. Krasnokutskaya (2011) also used deconvolution results to estimate auction models with unobserved heterogeneity. To our knowledge, however, our paper is the first application of (discrete) measurement error results to estimate an auction model where the number of potential bidders is unknown.

The issues considered in this paper are close to those considered in the literature on entry in auctions: eg. Li (2005), Li and Zheng (2006), Athey et al. (2005), Krasnokutskaya and Seim (2011), Haile et al. (2003). While the entry models considered in these papers differ, their one commonality is to model more explicitly bidders' participation decisions in auctions, which can cause the number of observed bidders A to differ from the number of potential bidders N^*. For instance, Haile et al. (2003) consider an endogenous participation model in which the number of potential bidders is observed by the researcher, and equal to the observed number of bidders (i.e., $N^* = A$), so that non-observability of N^* is not a problem. However, A is potentially endogenous, because it may be determined in part by auction-specific unobservables which also affect the bids. By contrast, in this paper we assume that N^* is unobserved, and that $N^* \neq A$, but we do not consider the possible endogeneity of N^*.[3]

4.2.1 Model

In this paper, we consider the case of first-price auctions under the symmetric independent private values (IPV) paradigm, for which identification and estimation are most transparent. For a thorough discussion of identification and estimation of these models when the number of potential bidders N^* is known, see Paarsch and Hong (2006), ch. 4. For concreteness, we focus on the case where a binding reserve price is the reason why the number of potential bidders N^* differs from the observed number of bidders, and is not known by the researcher.

There are N^* bidders in the auction, with each bidder drawing a private valuation from the distribution $F_{N^*}(x)$ which has support $[\underline{x}, \bar{x}]$. Furthermore, we assume the density of the private valuation $f_{N^*}(x)$ is bounded away from zero on $[\underline{x}, \bar{x}]$.[4] N^* can vary freely across the auctions, and while it is observed by the bidders, it is not known by the researcher. We allow the distribution of valuations $F_{N^*}(x)$ to vary across N^*.[5] There is a reserve price r, assumed to be fixed across all auctions, where $r > \underline{x}$.[6] The equilibrium bidding function

number of active bidders is *not observed* by bidders. Under her assumptions, however, the observed bids are i.i.d. samples from a homogeneous distribution, so that her estimation methodology would not work for the model considered in this paper.

[3] In principle, we recover the distribution of bids (and hence the distribution of valuations) separately for each value of N^*, which accommodates endogeneity in a general sense. However, because we do not model the entry process explicitly (as in the papers cited above), we do not deal with endogeneity in a direct manner.

[4] This assumption guarantees that the density of bids $g(b|N^*, b > r)$ is also bounded away from zero. See Guerre et al. (2000), Section 3.1 for detailed discussions.

[5] This is consistent with some models of endogenous entry. See section 4.2.6 below.

[6] Our estimation methodology can potentially also be used to handle the case where $N^*$is fixed across all auctions, but r varies freely across auctions.

for bidder i with valuation x_i is

$$b(x_i; N^*) \begin{cases} = x_i - \frac{\int_r^{x_i} F_{N^*}(s)^{N^*-1} ds}{F_{N^*}(x_i)^{N^*-1}} & \text{for } x_i \geq r \\ 0 & \text{for } x_i < r. \end{cases} \tag{4.7}$$

Hence, the number of bidders observed by the researcher is $A \equiv \sum_{i=1}^{N^*} \mathbf{1}(x_i > r)$, the number of bidders whose valuations exceed the reserve price.

For this case, the equilibrium bids are i.i.d. and, using the change-of-variables formula, the density of interest $g(b|N^*, b > r)$ is equal to

$$g(b|N^*, b > r) = \frac{1}{b'(\xi(b; N^*); N^*))} \frac{f_{N^*}(\xi(b; N^*))}{1 - F_{N^*}(r)}, \text{ for } b > r \tag{4.8}$$

where $\xi(b; N^*)$ denotes the inverse of the equilibrium bid function $b(\cdot; N^*)$ evaluated at b. In equilibrium, each observed bid from an N^*-bidder auction is an i.i.d. draw from the distribution given in Eq. (4.8), which does not depend on A, the observed number of bidders.

We propose a two-step estimation procedure. In the first step, the goal is to recover the density $g(b|N^*; b > r)$ of the equilibrium bids, for the truncated support $(r, +\infty)$. (For convenience, in what follows, we suppress the conditioning truncation event $b > r$.) To identify and estimate $g(b|N^*)$, we use the results from Hu (2008).

In second step, we use the methodology of Guerre et al. (2000) to recover the valuations x, from the density $g(b|N^*)$. For each b in the marginal support of $g(b|N^*)$, the corresponding valuation x is obtained by

$$\xi(b, N^*) = b + \frac{1}{N^* - 1} \left[\frac{G(b|N^*)}{g(b|N^*)} + \frac{F_{N^*}(r)}{1 - F_{N^*}(r)} \cdot \frac{1}{g(b|N^*)} \right]. \tag{4.9}$$

Notice that F_{N^*}, which is the valuation distributions, can also be recovered after we identify $g(b|N^*)$ for different N^*.

For most of this paper, we focus on the first step of this procedure, because the second step is a straightforward application of standard techniques.

4.2.2 Nonparametric Identification

In this section, we apply the results from Hu (2008) to show the identification of the first-price auction model with unknown N^*. The procedure requires two auxiliary variables:

1. a proxy N, e.g., the number of actual bidders A, which is a mismeasured version of N^*

2. an instrument Z, which could be a discretized second bid.

We observe a random sample of $\left\{\vec{b}_t, A_t\right\}$, where $\vec{b}_t$ denotes the vector of observed bids $\{b_{1t}, b_{2t}, \ldots, b_{A_t t}\}$. Note that we only observe A_t bids for each auction t. In what follows, we use b to denote a randomly chosen bid from each auction.

We assume the variables N, and N^* are both discrete, and they have the same support $\mathcal{N} = \{2..., K\}$ as the discretized second bid Z. Here K can be interpreted as the maximum number of bidders, which is fixed across all auctions.[7]

For convenience, we first define the following matrices which we shall use repeatedly. We use the notation $g(\cdots)$ to denote, generically, a probability mass or density function.

$$\begin{aligned} G_{b,N,Z} &\equiv [g(b, N = i, Z = j)]_{i,j}, \\ G_{N|N^*} &\equiv [g(N = i|N^* = k)]_{i,k}, \\ G_{N^*,Z} &\equiv [g(N^* = k, Z = j)]_{k,j}, \\ G_{N,Z} &\equiv [g(N = i, Z = j)]_{i,j}, \end{aligned}$$

and

$$G_{b|N^*} \equiv \begin{pmatrix} g(b|N^* = 2) & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & g(b|N^* = K) \end{pmatrix}. \tag{4.10}$$

All of these are $(K-1)$-dimensional square matrices.

The five conditions required for our identification argument are given here:

Assumption 4.2.1 $g(b|N^*, N, Z) = g(b|N^*)$.

Assumption 4.2.2 $g(N|N^*, Z) = g(N|N^*)$.

Assumption 4.2.3 $Rank\,(G_{N,Z}) = K - 1$.

Assumption 4.2.4 *For any $i, j \in \mathcal{N}$, the set $\{(b) : g(b|N^* = i) \neq g(b|N^* = j)\}$ has nonzero Lebesgue measure whenever $i \neq j$.*

Assumption 4.2.5 $N \leq N^*$.

In this section, we will show how Conditions 1-5 lead to the identification of the unknown elements $G_{b|N^*}$, $G_{N|N^*}$ and $G_{N^*,Z}$ (the former pointwise in b). The conditions will be discussed as they arise in the identification argument.

Condition 1 implies that N or Z affects the equilibrium density of bids only through the unknown number of potential bidders N^*. In the econometric literature, this is known as the "nondifferential" measurement error assumption. In what follows, we only consider values of b such that $g(b|N^*) > 0$, for $N^* = 2, \ldots, K$. This requires, implicitly, knowledge of the support of $g(b|N^*)$, which is typically unknown to the researcher. Below, when we discuss estimation, we present a two-step procedure to estimate $g(b|N^*)$ which circumvents this problem.

Condition 2 implies that the instrument Z affects the mismeasured N only through the number of potential bidders. Roughly, because N is a noisy measure of N^*, this condition requires that the noise is independent of the instrument Z, conditional on N^*.

[7] Our identification results still hold if Z has more possible values than N and N^*.

Examples of N and Z Before proceeding with the identification argument, we consider several examples of auxiliary variables (N, Z) which satisfy conditions 1 and 2.

1. One advantage to focusing on the IPV model is that A, the observed number of bidders, can be used in the role of N. Particularly, for a given N^*, the sampling density of any equilibrium bid exceeding the reserve price — as given in Eq. (4.8) above — does not depend on A, so that Condition 1 is satisfied.[8]

A good candidate for the instrument Z is a discretized second bid, and it depends on N^* through Eq.(4.7):

$$Z = b(N^*, x_z).$$

where x_z denotes the valuation of the bidder who submits the second bid Z. In order to satisfy conditions 1 and 2, we would require $b \perp Z|N^*$, and also $A \perp Z|N^*$, which are both satisfied in the IPV setting. The use of a second bid in the role of the instrument Z echoes the use of two bids per auction in the earlier identification results of Li et al. (2002) and Krasnokutskaya (2011). Hence, just as in those papers, our identification and estimation approach is applicable to any IPV auction with two or more bidders.

Because we are focused on the symmetric IPV model in this paper, we will consider this example in the remainder of this section, and also in our Monte Carlo experiments and in the empirical illustration.

2. A second possibility is that N is a noisy measure of N^*, as in example 2, but Z is an exogenous variable which directly determines participation:

$$\begin{aligned} N &= l(N^*, \upsilon) \\ N^* &= k(Z, \nu). \end{aligned} \tag{4.11}$$

In order to satisfy conditions 1 and 2, we would require $b \perp (\upsilon, Z)|N^*$, as well as $\upsilon \perp Z|N^*$. This implies that Z is excluded from the bidding strategy, and affects bids only through its effect on N^*.

Furthermore, in this example, in order for the second step of the estimation procedure (in which we recover bidders' valuations) to be valid, we also need to assume that $b \perp \nu|N^*$. Importantly, this rules out the case that the participation shock ν is a source of unobserved auction-specific heterogeneity.[9] Note that ν will generally be (unconditionally) correlated with the bids b, which our assumptions allow for. ■

By the law of total probability, the relationship between the observed distribution $g(b, N, Z)$ and the latent densities is as follows:

$$g(b, N, Z) = \sum_{N^*=2}^{K} g(b|N^*, N, Z)g(N|N^*, Z)g(N^*, Z). \tag{4.12}$$

[8] This is no longer true in affiliated value models.

[9] In the case when N^* is observed, correlation between bids and the participation shock ν can be accommodated, given additional restriction on the $k(\cdots)$ function. See Guerre et al. (2009) and Haile et al. (2003) for details. However, when N^* is unobserved, as is the case here, it is not clear how to generalize these results.

Under conditions 1 and 2, Eq. (4.12) becomes

$$g(b, N, Z) = \sum_{N^*=2}^{K} g(b|N^*)g(N|N^*)g(N^*, Z). \tag{4.13}$$

Eq. (4.13) can be written as

$$G_{b,N,Z} = G_{N|N^*}G_{b|N^*}G_{N^*,Z}. \tag{4.14}$$

Condition 4.2.2 implies that

$$g(N, Z) = \sum_{N^*=2}^{K} g(N|N^*)g(N^*, Z), \tag{4.15}$$

which, using the matrix notation above, is equivalent to

$$G_{N,Z} = G_{N|N^*}G_{N^*,Z}. \tag{4.16}$$

Equations (4.14) and (4.16) summarize the unknowns in the model, and the information in the data. The matrices on the left-hand sides of these equations are quantities which can be recovered from the data, whereas the matrices on the right-hand side are the unknown quantities of interest. As a counting exercise, we see that the matrices $G_{b,N,Z}$ and $G_{N,Z}$ contain $2(K-1)^2 - (K-1)$ known elements, while the unknown matrices $G_{N|N^*}$, $G_{N^*,Z}$ and $G_{b|N^*}$ contain at most a total of also $2(K-1)^2 - (K-1)$ unknown elements. Hence, in principle, there is enough information in the data to identify the unknown matrices. The key part of the proof below is to characterize the solution and give conditions for uniqueness. Moreover, the proof is constructive in that it immediately suggests a way for estimation.

Eq. (4.16) implies that

$$Rank\left(G_{N,Z}\right) \leq \min\left\{Rank\left(G_{N|N^*}\right), Rank\left(G_{N^*,Z}\right)\right\}. \tag{4.17}$$

Hence, it follows from Condition 3 that $Rank\left(G_{N|N^*}\right) = K-1$ and $Rank\left(G_{N^*,Z}\right) = K-1$. In other words, the matrices $G_{N,Z}$, $G_{N|N^*}$, and $G_{N^*,Z}$ are all invertible.[10] Therefore, postmultiplying both sides of Eq. (4.14) by $G_{N,Z}^{-1} = G_{N^*,Z}^{-1}G_{N|N^*}^{-1}$, we obtain the key equation

$$G_{b,N,Z}G_{N,Z}^{-1} = G_{N|N^*}G_{b|N^*}G_{N|N^*}^{-1}. \tag{4.18}$$

The matrix on the left-hand side can be formed from the data. For the expression on the right-hand side, note that because $G_{b|N^*}$ is diagonal (cf. Eq. (4.10)), the RHS matrix represents an eigenvalue-eigenvector decomposition of the LHS matrix, with $G_{b|N^*}$ being the diagonal matrix of eigenvalues, and $G_{N|N^*}$ being the corresponding matrix of eigenvectors. This is the key representation which will identify and facilitate estimation of the unknown

[10]Note that Condition 3 is directly testable from the sample. It essentially ensures that the instrument Z affects the distribution of the proxy variable N (resembling the standard instrumental relevance assumption in usual IV models).

matrices $G_{N|N^*}$ and $_{b|N^*}$.

In order to make the eigenvalue-eigenvector decomposition in Eq. (4.18) unique, Condition 4 is required. This condition, which is actually implied by equilibrium bidding, guarantees that the eigenvalues in $G_{b|N^*}$ are distinctive for some bid b, which ensures that the eigenvalue decomposition in Eq. (4.18) exists and is unique, for some bid b. Moreover, it guarantees that all the linearly independent eigenvectors are identified from the decomposition in Eq. (4.18).[11]

Given Condition 4, Eq. (4.18) shows that an eigenvalue decomposition of the observed $G_{b,N,Z}G_{N,Z}^{-1}$ matrix identifies $G_{b|N^*}$ and $G_{N|N^*}$ up to a normalization and ordering of the columns of the eigenvector matrix $G_{N|N^*}$.

There is a clear appropriate choice for the normalization constant of the eigenvectors; because each column of $G_{N|N^*}$ should add up to one, we can multiply each element $G_{N|N^*}(i,j)$ by the reciprocal of the column sum $\sum_i G_{N|N^*}(i,j)$, as long as $G_{N|N^*}(i,j)$ is non-negative.

The appropriate ordering of the columns of $G_{N|N^*}$ is less clear, and in order to complete the identification, we need an additional condition which pins down the ordering of these columns. Condition 5, which posits that $N \leq N^*$ is one example of such an ordering condition. It is natural, and automatically satisfied, when $N = A$, the observed number of bidders. This condition implies that for any $i, j \in \mathcal{N}$

$$g\left(N = j | N^* = i\right) = 0 \text{ for } j > i. \tag{4.19}$$

In other words, $G_{N|N^*}$ is an upper-triangular matrix. Since the triangular matrix $G_{N|N^*}$ must be invertible (by Eq. (4.17), its diagonal entries are all nonzero, i.e.,

$$g\left(N = i | N^* = i\right) > 0 \text{ for all } i \in \mathcal{N}. \tag{4.20}$$

In other words, Condition 5 implies that, once we have the columns of $G_{N|N^*}$ obtained as the eigenvectors from the matrix decomposition (4.18), the right ordering can be obtained by re-arranging these columns so that they form an upper-triangular matrix.

Hence, the arguments in this section have shown the following result:

Theorem 4.2.1 *Under Conditions 4.2.1-4.2.5, $G_{b|N^*}$, $G_{N|N^*}$ and $G_{N^*,Z}$ are identified (the former pointwise in b).*

4.2.3 Nonparametric Estimation: Two-step Procedure

In this section, we give details on the estimation of $(b|N^*)$ given observations of (b, N, Z), for the symmetric independent private values model. In the key equation (4.18), the matrix

[11] Specifically, suppose that for some value $\widetilde{b}$, $g(\widetilde{b}|N^* = i) = g(\widetilde{b}|N^* = j)$, which implies that the two eigenvalues corresponding to $N^* = i$ and $N^* = j$ are the same. In this case, the two corresponding eigenvectors cannot be uniquely identified, because any linear combination of the two eigenvectors is still an eigenvector. Condition 4.2.4 guarantees that there exists another value $\overline{b}$ such that $g(\overline{b}|N^* = i) \neq g(\overline{b}|N^* = j)$. Because Eq. (4.18) holds for every b, implying that $g(\widetilde{b}|N^* = i)$ and $g(\overline{b}|N^* = i)$ correspond to the same eigenvector, as do $g(\widetilde{b}|N^* = j)$ and $g(\overline{b}|N^* = j)$, we can use the value $\overline{b}$ to identify the two eigenvectors corresponding to $N^* = i$ and $N^* = j$.

$G_{N|N^*}$ is identical for all b.[12] This suggests a convenient two-step procedure for estimating the unknown matrices $G_{N|N^*}$ and $G(b|N^*)$.

Step One In Step 1, we estimate the eigenvector matrix $G_{N|N^*}$. To maximize the convergence rate in estimating $G_{N|N^*}$, we average across values of the bid b. Specifically, from Eq. (4.13), we have

$$E(b|N,Z)g(N,Z) = \sum_{N^*=2}^{K} E(b|N^*)g(N|N^*)g(N^*,Z) \tag{4.21}$$

where $E[\cdot|\cdot]$ denote conditional expectation. Define the matrices

$$G_{Eb,N,Z} \equiv \left[E\left(b|N=i,Z=j\right)g(N=i,Z=j)\right]_{i,j}, \tag{4.22}$$

and

$$G_{Eb|N^*} \equiv \begin{pmatrix} E\left[b|N^*=2\right] & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & E\left[b|N^*=K\right] \end{pmatrix}.$$

Then

$$G_{Eb,N,Z} = G_{N|N^*}G_{Eb|N^*}G_{N^*,Z}$$

and, as before, postmultiplying both sides of this equation by $G_{N,Z}^{-1} = G_{N^*,Z}^{-1}G_{N|N^*}^{-1}$, we obtain an integrated version of the key equation:

$$G_{Eb,N,Z}G_{N,Z}^{-1} = G_{N|N^*}G_{Eb|N^*}G_{N|N^*}^{-1}. \tag{4.23}$$

This implies

$$G_{N|N^*} = \psi\left(G_{Eb,N,Z}G_{N,Z}^{-1}\right),$$

where $\psi(\cdot)$ denotes the mapping from a square matrix to its eigenvector matrix following the identification procedure in the previous section.[13] As mentioned in Hu (2008), the function $\psi(\cdot)$ is a nonstochastic analytic function. Therefore,we may estimate $G_{N|N^*}$ as follows:

$$\widehat{G}_{N|N^*} := \psi\left(\widehat{G}_{Eb,N,Z}\widehat{G}_{N,Z}^{-1}\right), \tag{4.24}$$

[12]This also implies that there is a large degree of overidentification in this model, and suggests the possibility of achieving identification with weaker assumptions. In particular, it may be possible to relax the non-differentiability condition 1 so that we require $g(b|N^*,N,Z) = g(b|N^*)$ only at one particular value of b. We are exploring the usefulness of such possibilities in ongoing work.

[13]In order for $G_{N|N^*}$ to be recovered from this eigenvector decomposition, Condition 4 from the previous section must be strengthened so that the conditional means $E[b|N^*]$, which are the eigenvalues from this decomposition, are distinct for every N^*.

where $\widehat{G}_{Eb,N,Z}$ and $\widehat{G}_{N,Z}$ may be constructed directly from the sample. In our empirical example, we estimate $\widehat{G}_{Eb,N,Z}$ using a sample average:

$$\widehat{G}_{Eb,N,Z} = \left[\frac{1}{T} \sum_t \frac{1}{N_j} \sum_{i=1}^{N_j} b_{it} \mathbf{1}(N_t = N_j, Z_t = Z_k) \right]_{j,k}. \tag{4.25}$$

Step Two In Step 2, we estimate $g(b|N^*)$. With $G_{N|N^*}$ estimated by $\widehat{G}_{N|N^*}$ in step 1, we may proceed to estimate $g(b|N^*)$, pointwise in b. First, consider

$$g(b, N) = \sum_{N^*} g(N|N^*) g(b, N^*)$$

which, in matrix form, is

$$\overrightarrow{g}(b, N) = G_{N|N^*} \overrightarrow{g}(b, N^*),$$

where the vector of densities $\overrightarrow{g}(\cdot, \cdot) \equiv \left[g(b, N = 2), g(b, N = 3), \ldots, g(b, N = K) \right]^T$.

Define $e_{N^*} = (0, ..., 0, 1, 0, ..., 0)^T$, where 1 is at the N^*-th position in the vector. This relation suggests that we may estimate the joint density $g(b, N^*)$ as follows:

$$\widehat{g}(b, N^*) = e_{N^*}^T \widehat{G}_{N|N^*}^{-1} \overrightarrow{\widehat{g}}(b, N),$$

where $\widehat{G}_{N|N^*}$ is estimated in step 1, and we use a kernel estimate for each element of the vector $\overrightarrow{\widehat{g}}(b, N) = \left[\widehat{g}(b, N = 2), \widehat{g}(b, N = 3), \ldots, \widehat{g}(b, N = K) \right]^T$:

$$\widehat{g}(b, N_j) = \left[\frac{1}{Th} \sum_t \frac{1}{N_t} \sum_{i=1}^{N_t} K\left(\frac{b - b_{it}}{h} \right) \mathbf{1}(N_t = N_j) \right]. \tag{4.26}$$

Given this estimate of $\widehat{g}(b, N^*)$, it is straightforward to estimate $g(b|N^*)$. Define $\overrightarrow{g}_N$, and $\overrightarrow{g}_{N^*}$ as the vectors of distributions for N and N^*, respectively.[14] Then,

$$\overrightarrow{g}_N = G_{N|N^*} \overrightarrow{g}_{N^*}.$$

We may then estimate

$$\widehat{\Pr(N^*)} = e_{N^*}^T \widehat{G}_{N|N^*}^{-1} \overrightarrow{\widehat{g}}(N),$$

where $\overrightarrow{\widehat{g}}(N) \equiv \left[\frac{1}{T} \sum_t \mathbf{1}_{N_t=2}, \ldots, \frac{1}{T} \sum_t \mathbf{1}_{N_t=K} \right]$ can be recovered directly from the sample. Therefore, the conditional bid densities $g(b|N^*)$ may be estimated as

$$\widehat{g}(b|N^*) = \frac{e_{N^*}^T \widehat{G}_{N|N^*}^{-1} \overrightarrow{\widehat{g}}(b, N)}{e_{N^*}^T \widehat{G}_{N|N^*}^{-1} \overrightarrow{\widehat{g}}(N)}. \tag{4.27}$$

Analogously, we can also recover $F(b|N^*)$, the empirical conditional CDF's for the bids,

[14] For example, if $N^* = \{2, 3, 4\}$, then $\overrightarrow{g}_{N^*} = \left\{ \Pr(N^* = 2), \Pr(N^* = 3), \Pr(N^* = 4) \right\}^T$.

using the conditional empirical CDF:

$$\widehat{F}(b|N^*) = \frac{e_{N^*}^T \widehat{G}_{N|N^*}^{-1} \overrightarrow{\widehat{F}}(b, N)}{e_{N^*}^T \widehat{G}_{N|N^*}^{-1} \overrightarrow{\widehat{g}}(N)}, \tag{4.28}$$

where $\widehat{F}(b, N)$ denotes the vector of empirical CDF's with elements:

$$\widehat{F}(b, N_j) = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbf{1}(b_{it} < b, N_t = N_j), \quad N_j = 2, \ldots, K \tag{4.29}$$

which can be recovered from the sample.

In the Monte Carlo experiments and empirical application, we estimated both bid CDF's (using Eq. (4.28)) and bid densities (using Eq. (4.27)) to assess the performance of our estimation procedure. An advantage of empirical CDFs over kernel density estimates is that we do not need to worry about the effects of bandwidth choice on the performance of our estimator.

Because $\Pr(N^* = K|A = K) = 1$, and $G_{N|N^*}$ is an upper-triangular matrix, our estimates of $F(b|N^* = K)$ and $g(b|N^* = K)$ are identical to, respectively, $F(b|A = K)$ and $g(b|A = K)$. Our estimation requires a value for K, the upper bound for the number of potential bidders. In practice, K is unknown, but we set it to be the maximum number of observed bidders, which is a super-consistent estimate.[15]

The bid b may have a different unknown support for different N^*. That is,

$$g(b|N^*) = \begin{cases} > 0 & \text{for } b \in [r, u_{N^*}] \\ = 0 & \text{otherwise} \end{cases},$$

where u_{N^*}, the upper bound of the support of $g(b|N^*)$, may not be known by the researcher. In practice, we estimate the upper bound u_{N^*} as follows:

$$\hat{u}_{N^*} = \sup \{b : \hat{g}(b|N^*) > 0\}.$$

In general, using the supremum to estimate the upper bound of an observed random sample is somewhat naïve. Estimation of the support of an observed random sample has been extensively studied in the statistics literature (see Cuevas and Rodríguez-Casal (2004) for i.i.d. data, and Delaigle and Gijbels (2006a, 2006b) for data measured with error), and our estimate of u_{N^*} can be improved by employing these methods. However, because an unbiased and consistent estimator of u_{N^*} is all we need, the naïve estimator $\widehat{u}_{N^*}$ is sufficient for our purposes, and we do not consider more sophisticated estimators in this paper.[16]

The asymptotic properties of our estimator are analyzed in detail in the appendix. Here

[15] This is obvious if the reserve price is zero. However, this is also valid when the reserve price is greater than zero because, even when $r > 0$, the probability that the observed number of bidders is equal to K is still strictly positive.

[16] This naïve estimator for the upper bound of the support of bids is commonly used in the auction literature, e.g., see Donald and Paarsch (1993) and Guerre et al. (2000), among others.

we provide a brief summary. Given the discreteness of N, Z, and the use of a sample average to construct $\widehat{G}_{Eb,N,Z}$ (via. Eq. (4.25)), the estimates of $\widehat{G}_{N|N^*}$ (obtained using Eq. (4.24)) and $\widehat{G}_{N,Z}$ should converge at a $\sqrt{T}$-rate (where T denotes the total number of auctions).

Hence, pointwise in b, the convergence properties of $\widehat{g}(b|N^*)$ to $g(b|N^*)$, where $\widehat{g}(b|N^*)$ is estimated using Eq. (4.27), will be determined by the convergence properties of the kernel estimate of $g(b,N)$ in Eq. (4.26), which converges at a rate slower than $\sqrt{T}$. In the Appendix, we show that, pointwise in b, $(Th)^{1/2}\left[\widehat{g}(b|N^*) - g(b|N^*)\right]$ converges to a normal distribution. We also present a uniform convergence rate for $\widehat{g}(b|N^*)$. As for the empirical distribution $\widehat{F}(b|N^*)$, it is well known that $T^{1/2}[\widehat{F}(b,N) - F(b,N)]$ converges to a normal distribution with mean zero. Because $\widehat{G}_{N|N^*}$ converges at a $\sqrt{T}$-rate, $\widehat{F}(b|N^*)$ also converges at $\sqrt{T}$-rate. We omit the proof of this as the argument is similar to the proof for $\widehat{g}(b|N^*)$.

The matrix $G_{N|N^*}$, which is a by-product of the estimation procedure, can be useful for specification testing, when $N = A$, the observed number of bidders. In the scenario where the difference between the observed number of bidders A and the number of potential bidders N^* arises from a binding reserve price, and that the reserve price r is fixed across all the auctions with the same N^* in the dataset, it is well-known (cf. Paarsch (1997)) that

$$A|N^* \;\sim\; \text{Binomial}(N^*, 1 - F_{N^*}(r)) \tag{4.30}$$

where $F_{N^*}(r)$ denotes the CDF of bidders' valuations in auctions with N^* potential bidders, evaluated at the reserve price. This suggests that the recovered matrix $G_{A|N^*}$ can be useful in two respects. First, using Eq. (4.30), the truncation probability $F_{N^*}(r)$ could be estimated, for each value of N^*. This is useful when we use the first-order condition (4.9) to recover bidders' valuations. Alternatively, we could also test whether the columns of $G_{A|N^*}$, which correspond to the probabilities $\Pr(A|N^*)$ for a fixed N^*, are consistent with the binomial distribution in Eq. (4.30).

4.2.4 Monte Carlo Evidence

In this section, we present some Monte Carlo evidence for our estimation procedure. We consider first price auctions where bidders' valuations $x_i \sim U[0,1]$, independently across bidders i. With a reserve price $r > 0$, the equilibrium bidding strategy with N^* bidders is:

$$b^*(x;N^*) = \mathbf{1}_{x \geq r}\left\{\left(\frac{N^*-1}{N^*}\right)x + \frac{1}{N^*}\left(\frac{r}{x}\right)^{N^*-1} r\right\} \tag{4.31}$$

For each auction t, we generate the equilibrium bids b_{jt}, for $j = 1, \ldots N_t^*$, as well as (N_t^*, N_t, Z_t). The proxy N_t is taken to be the number of observed bidders A_t, and Z_t is a discretized second bid. The number of potential bidders N_t^* for each auction t is generated uniformly on $\{2, 3, \ldots, K\}$, where K, the maximum number of bidders, is set at 4. For each auction t, and each bidder $j = 1, \ldots, N_t^*$, we draw valuations $x_j \sim U[0,1]$, and construct the corresponding equilibrium bids using Eq.(4.31). Subsequently, the number of observed bidders is determined as the number of bidders whose valuations exceed the reserve price:

$A_t = \sum_{j \in \mathcal{N}_t^*} \mathbf{1}(x_j \geq r)$.

The estimation procedure in section 4.2.3 requires $A_t \geq 2$ for each t, so that the supports of A_t and N_t^* coincide. For this reason, we discard all the auctions with $A_t = 1$;[17] for each of the remaining auctions, we randomly pick a pair of bids (b_{1t}, b_{2t}), and use a discretized version of the second bid b_{2t} in the role of Z_t.[18]

Results

We present results from $S = 400$ replications of a simulation experiment. The performance of our estimation procedure is illustrated in Figures 4.1 and 4.4. The estimator performs well for all values of $N^* = 2, 3, 4$, and for modest-sized datasets of $T = 1000$ and $T = 400$ auctions, especially for the empirical bid distribution functions. Across the Monte Carlo replications, the estimated distribution and density functions track the actual densities quite closely. In these graphs, we also plot the bid CDF's (labelled"$G(b|A)$") and densities ($g(b|A)$) conditional on A, which are "naïve" estimators for $F(b|N^*)$, and $g(b|N^*)$, respectively. For $N^* = 2, 3$, our estimator outperforms the naïve estimator, especially for the case of $N^* = 2$. As we mentioned earlier, for $N^* = 4$, our estimates coincide with the naïve estimates.

In Figure 4.3 and 4.6, we present estimates of bidders' valuations. In each graph on the left-hand-side of the figure, we graph the bids against three measures of the corresponding valuation: (i) the actual valuation, computed from Eq. (4.9) using the actual bid densities $g(b|N^*)$, and labeled "True values"; (ii) the estimated valuations using our estimates of $g(b|N^*)$, labeled "Estimated value"[19]; and (iii) naïve estimates of the values, computed using $g(b|A)$, the observed bid densities conditional on the observed number of bidders.[20]

The graphs show that there are sizable differences between the value estimates, across all values of the bids. For all values of N^*, we see that our estimator tracks the true values quite closely. In contrast, the naïve approach underestimates the valuations. This is to be expected — because $N^* \geq A$, the set of auctions with a given value of A actually have a true level of competition larger than A. Hence, the naïve approach overstates the true level of competition, which leads to underestimation of bidders' markdowns $(x - b)/x$. The markdowns implied by our valuation estimates are shown in the right-hand-side graphs in Figure 4.3 and Figure 4.6.

4.2.5 Empirical Illustration

In this section, we illustrate our methodology using a dataset of low-bid construction procurement auctions held by the New Jersey Department of Transportation (NJDOT) in the

[17] Because of Condition 1, ignoring the auctions with $A_t = 1$ does not affect the consistency of the estimates of the bid distributions $g(b|N^*)$. There is only an efficiency impact from using fewer observations.

[18] Specifically, in this experiment, bids are distributed on $[0.3, 0.75]$, and both $N^*, A \in \{2, 3, 4\}$. Hence, the discretized second bid Z_t also takes values $\{2, 3, 4\}$ as follows: if $b_{2t} \in [0.3, 0.55], Z = 2$; $b_{2t} \in [0.55, 0.675], Z = 3$; $b_{2t} \in [0.675, 0.55], Z = 4$.

[19] In computing these valuations, the truncation probability $F(r)$ in Eq. (4.9) is obtained from the first-step estimates of the misclassification probability matrix $G_{N|N^*}$ as $\hat{F}(r) = 1 - \left[\hat{G}(N^*|N^*)\right]^{1/N^*}$.

[20] In computing the values for the naïve approach, we use the first-order condition $\xi(b; A) = b + \frac{G(b|A)}{(A-1)\cdot g(b|A)}$, which ignores the possibility of a binding reserve price.

Figure 4.1: Estimates of bid distribution functions and densities: $K = 4$, $T = 1000$

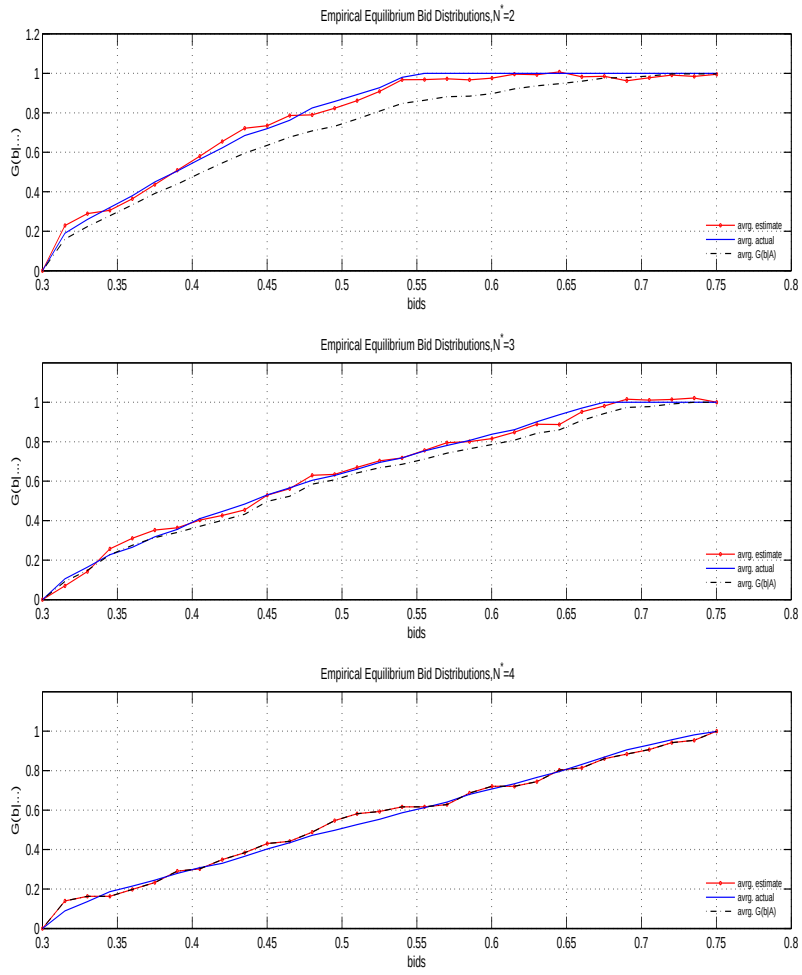


Figure 4.2: Estimates of bid distribution functions and densities: $K = 4$, $T = 1000$

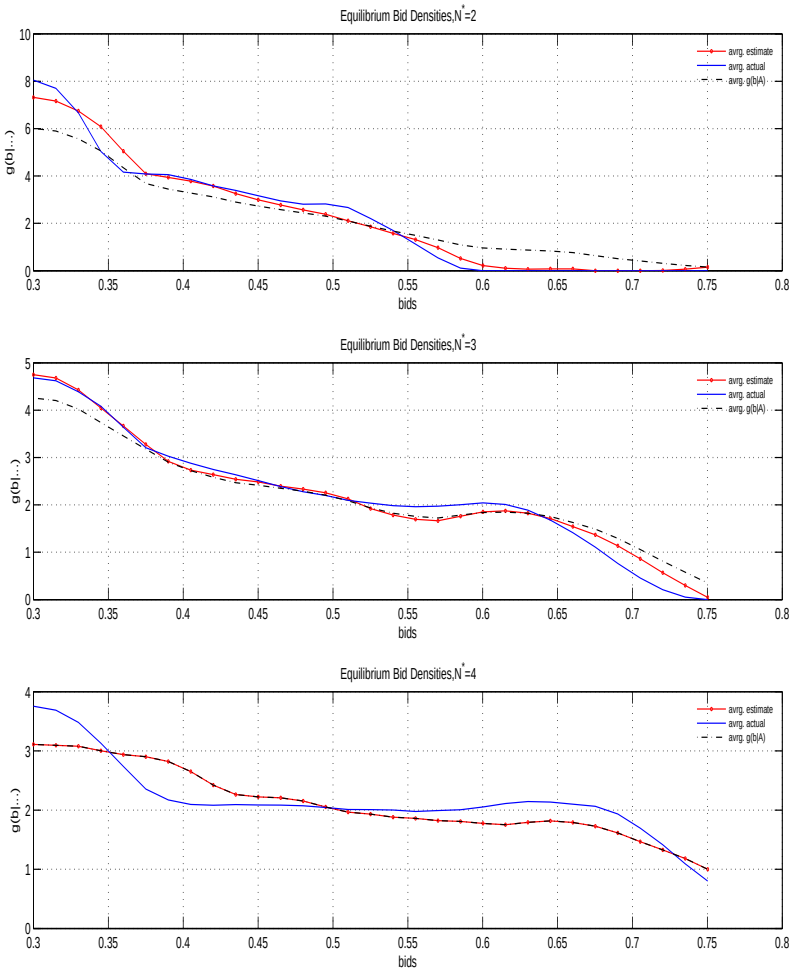


Figure 4.3: Estimates of bid functions and implied markdowns, $K = 4, T = 1000$

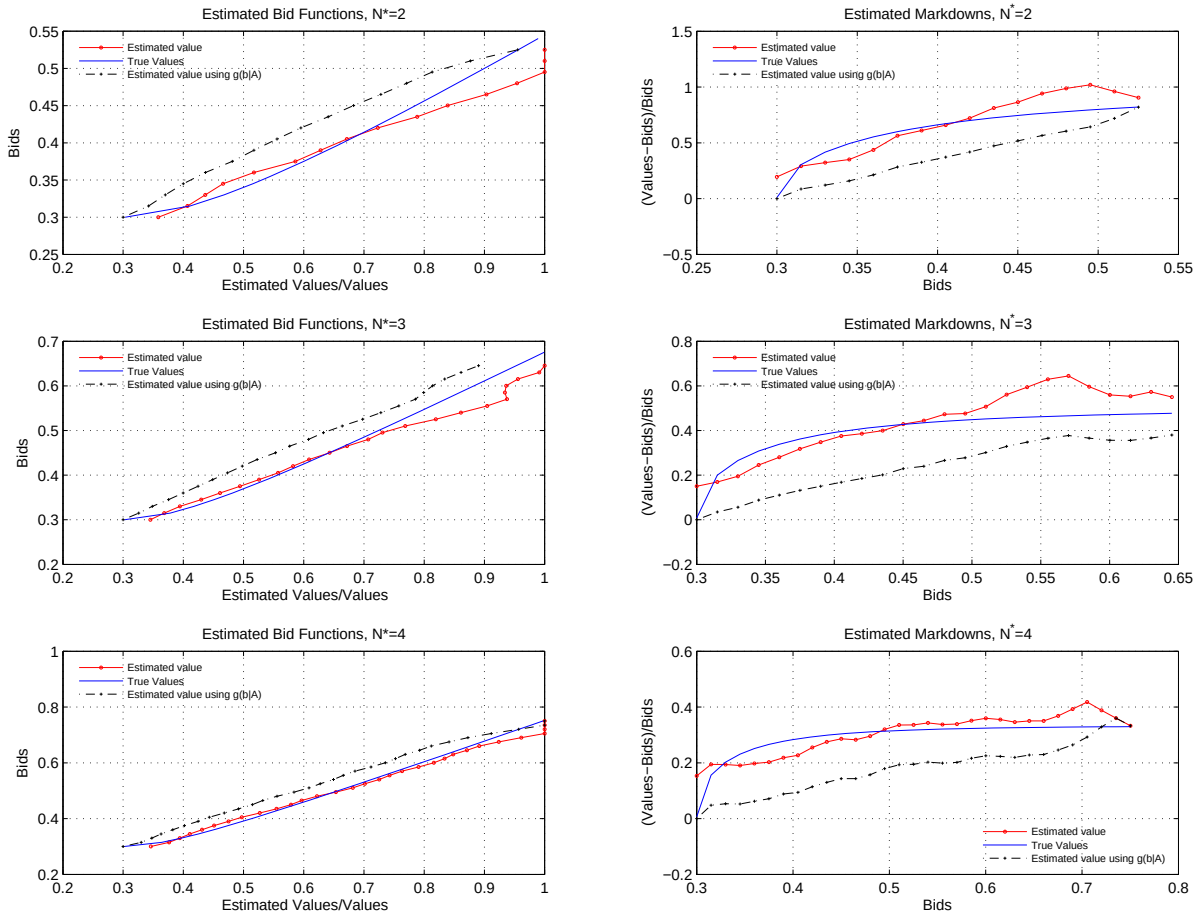


Figure 4.4: Estimates of bid distribution functions and densities: $K = 4$, $T = 400$

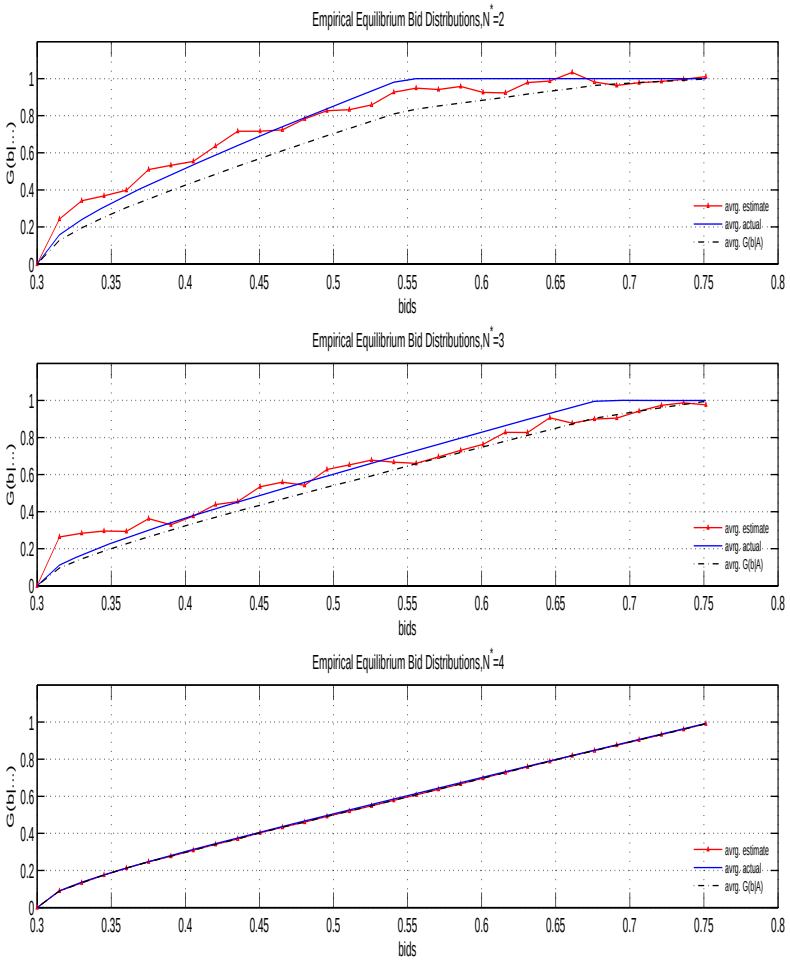


Figure 4.5: Estimates of bid distribution functions and densities: $K = 4$, $T = 400$

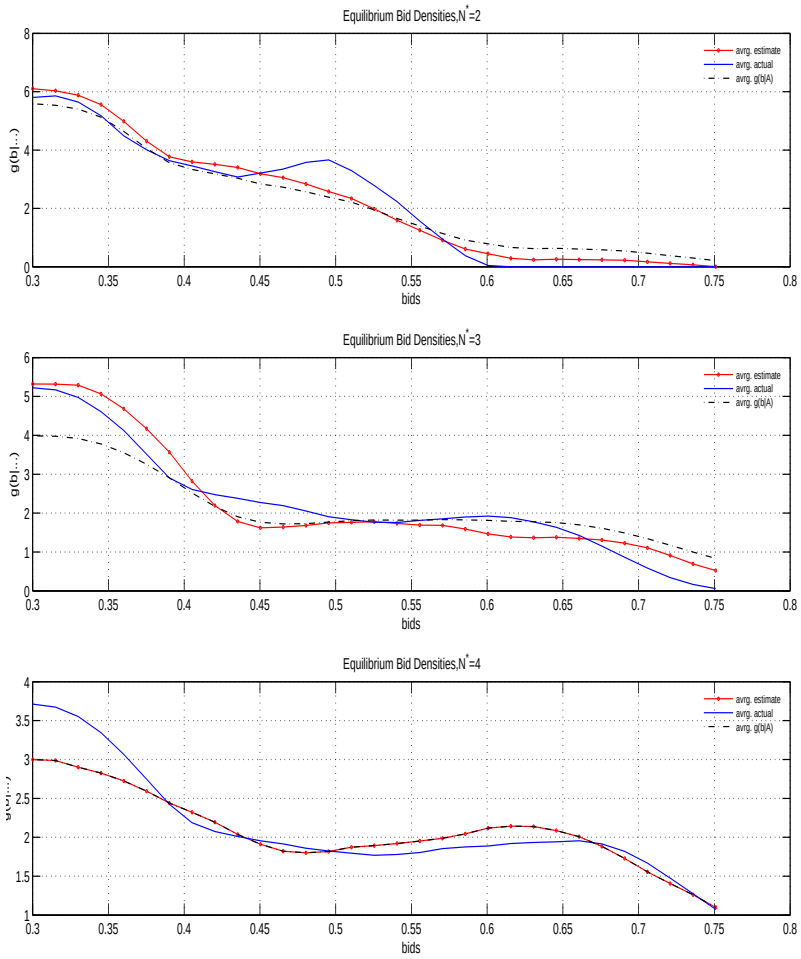
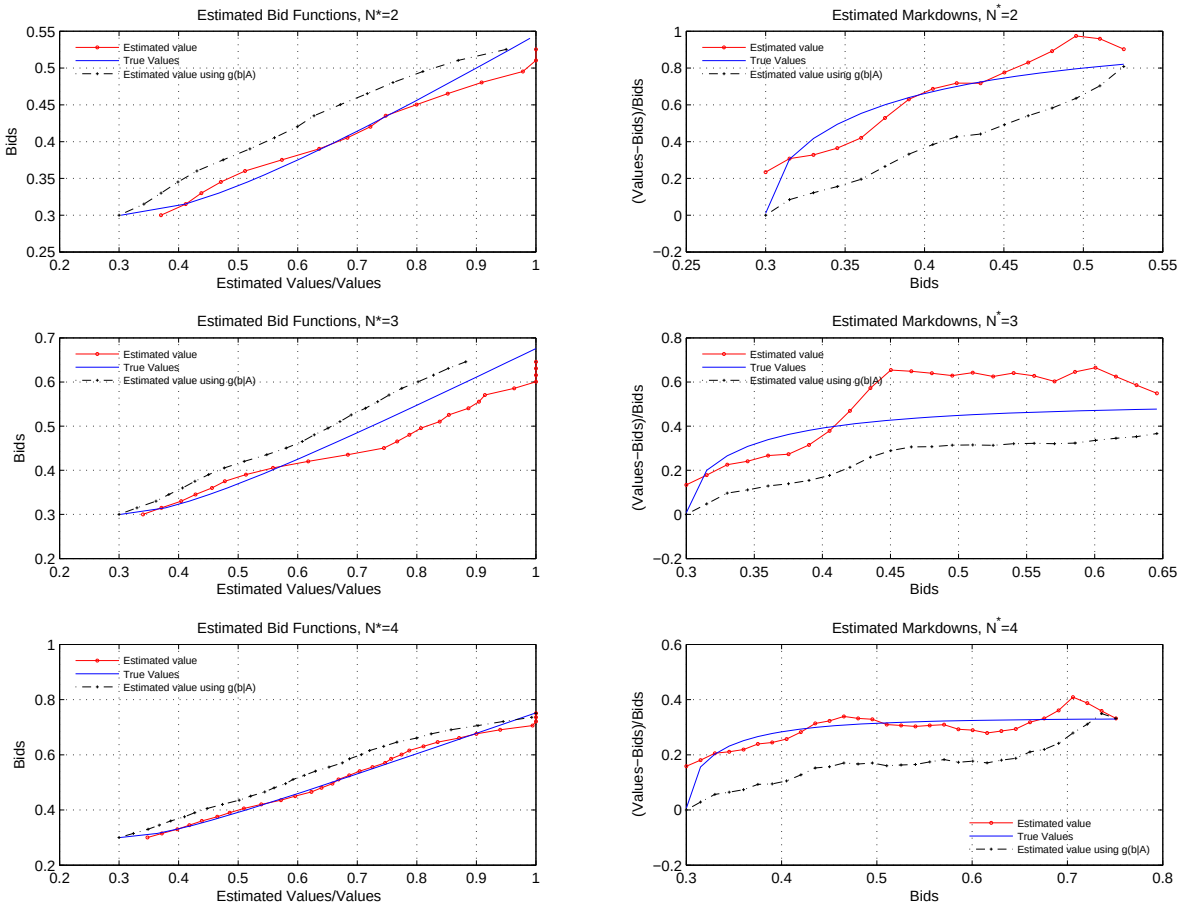


Figure 4.6: Estimates of bid functions and implied markdowns, $K = 4, T = 400$



years 1989–1997. This dataset was previously analyzed in Hong and Shum (2002), and a full description of it is given there. Moreover, Hong and Shum's (2002) analysis allows for common values, whereas we just have a simpler IPV model in the application here.[21]

Among all the auctions in our dataset, we focus on highway work construction projects, for which the number of auctions is the largest. In Table 4.1, we present some summary statistics on the auctions used in the analysis. Note that there were six auctions with just one bidder, in which non-infinite bids were submitted. If the observed number of bidders A is equal to N^*, the number of potential bidders observed by bidders when they bid, then the non-infinite bids observed in these one-bidder auctions is difficult to explain from a competitive bidding point of view.[22] However, occurrences of one-bidder auctions is a sign that the observed number of bidders is less than the potential number of bidders, perhaps due to an implicit reserve price. The methodology developed in this paper allows for this possibility.

For the two auxiliary variables, we used A, the number of observed bidders, in the role of the noisy measure N. We only analyze auctions with $A = 2, 3, 4$. Correspondingly, N^* also takes three distinct values from $\{2, 3, 4\}$. Because we focus on this range of small A, we assume that all the auctions are homogeneous.[23] In the role of the instrument Z, we use a second bid, discretized to take three values, so that the support of Z is the same as that of A.[24]

Furthermore, we use the ordering condition 5, which implies $A \leq N^*$, which is consistent with the story that bidders decide not to submit a bid due to an implicit reserve price. By an implicit reserve price, we mean a reserve price that bidders observe at the time of bidding, while not the econometrician. While there was no explicit reserve price in these auctions, there may have been an implicit reserve price, which can be understood as bidders' common beliefs regarding the upper bound of bids that the auctioneer is willing to consider.[25]

Because we model these auctions in a simplified setting, we do not attempt a full analysis of these auctions. Rather, this exercise highlights some practical issues in implementing the

[21] We are uncertain how to extend our estimation approach to common (or affiliated) value settings, and are exploring this in ongoing work.

[22] Indeed, (Li and Zheng, 2006, pg. 9) point out that even when bidders are uncertain about the number of competitors they are facing, finite bids cannot be explained when bidders face a non-zero probability that they could be the only bidder.

[23] We also considered an alternative specification where we control for observed auction-specific heterogeneity via preliminary regressions of bids on auction characteristics, and then perform the analysis using the residuals from these regressions. The resulting estimates of the bid distributions (available from the authors upon request) were qualitatively similar to, but noisier than the results presented here. This may be due to the weak correlation between the residuals and N^*. Our identification scheme relies critically on the correlation between bids and N^*, and if the auction characteristics were strongly related with, and affect the bids through N^*, using the residuals from the regressions in place of the bids may eliminate much of the correlation, leading to noisier estimates.

[24] Namely, we set $Z_t = 2$ if the second bid b_t is less than the 25th percentile of all the second bids; between the 25th and the 75th percentile, $Z_t = 3$; greater that the 75th, $Z_t = 4$. We tried several other alternatives, to ensure that the results are robust. In general, even if the support of Z exceeds that of A, the rank of $G_{A,Z}$ remains the same, but the model is overidentified in the sense that there are more instruments than needed. Our estimation approach can be extended to this case by using the generalized inverse of $G_{A,Z}$, but we did not pursue this possibility here.

[25] In conversations with a NJDOT authority, we were told that bids which were deemed excessive could be rejected outright at the discretion of the auction officials, which is consistent with an implicit reserve price.

Table 4.1: Summary statistics of procurement auction data
Highway work auctions, worktype=4:
Only auctions with $A = 2, 3, 4$ were used in empirical analysis

Observed # bidders (A)	# aucs.	Freq.	avg. bid[a]
1	6	2.96	0.575
2	11	5.42	1.495
3	31	15.27	1.692
4	46	22.67	1.843
5+	109	53.69	4.034

[a]: in millions of 1989$

estimation methodology. There are three important issues. First, the assumption that $A \leq N^*$ implies that the matrix on the right-hand side of the key equation (4.23) should be upper triangular, and hence that the matrix on the left-hand side, $G_{Eb,N,Z}G_{N,Z}^{-1}$, which is observed from the data, should also be upper-triangular. In practice, this matrix may not be upper-triangular. However, we do not impose upper-triangularity on $G_{Eb,N,Z}G_{N,Z}^{-1}$ in the first step of estimation. Instead, we constrain the estimated matrix $\widehat{G}_{A|N^*}$ to be upper-triangular in the second step of estimation. Doing so has no effect on the asymptotic consistency and convergence properties of $\widehat{G}_{A|N^*}$ since $G_{Eb,N,Z}G_{N,Z}^{-1}$ is upper-triangular asymptotically, i.e., with probability 1, the lower-triangular elements of $G_{Eb,N,Z}G_{N,Z}^{-1}$ vanish. [26]

Second, even after imposing upper-triangularity on estimated $G_{A|N^*}$, it is still possible that the eigenvectors and eigenvalues could have negative elements, which is inconsistent with the interpretation of them as densities and probabilities.[27] When our estimate of the densities $g(b|N^*)$ took on negative values, our remedy was to set the density equal to zero, but normalize our density estimate so that the resulting density integrated to one.[28]

Third, for low-bid procurement auctions, the optimal bidding strategy, analogous to Eq. (4.7) above, is:

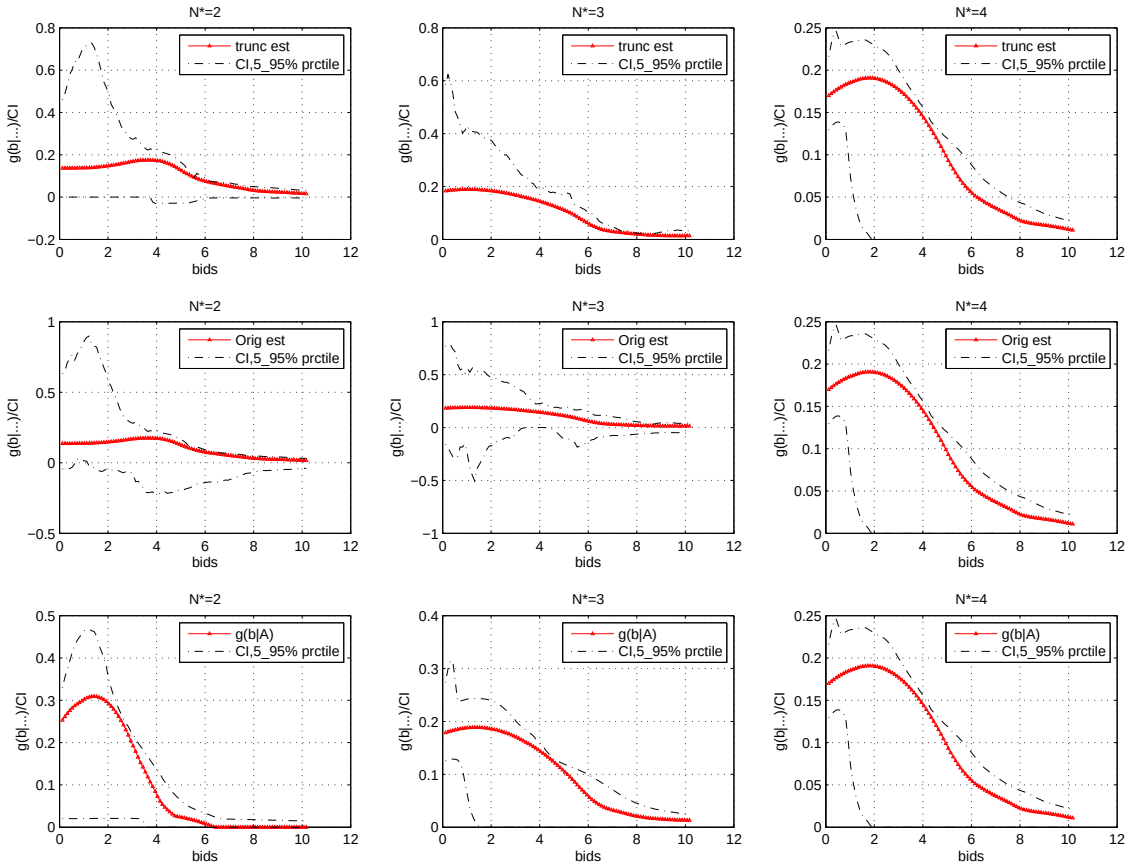
$$b(x_i; N^*) = \begin{cases} x_i + \frac{\int_{x_i}^{r}(1-F_{N^*}(s))^{N^*-1}ds}{(1-F_{N^*}(x_i))^{N^*-1}} & \text{for } x_i \leq r; \\ 0 & \text{for } x_i > r. \end{cases} \tag{4.32}$$

[26]Indeed, in the Monte Carlo simulations, we sometimes also had to impose this on the simulated data, as the $G_{Eb,N,Z}G_{N,Z}^{-1}$ matrix could be non-upper triangular due to small sample noise. In a previous version of the paper, we imposed upper-triangularity directly on $G_{Eb,N,Z}G_{N,Z}^{-1}$. Both methods have no effect on the asymptotic consistency and convergence properties on our estimator, but clearly the method in current version is more plausible since we did not impose any restriction on data-driven matrix $G_{Eb,N,Z}G_{N,Z}^{-1}$.

[27]This issue also arose in our Monte Carlo studies, but went away when we increased the sample size.

[28]Here we follow the recommendation of Efromovich (1999), pg. 63. This remedy does not affect the asymptotic properties of our estimator in that asymptotically $g(b|N^*)$ is bounded away from zero on its support, as we mentioned in footnote 4.

Figure 4.7: Highway work projects, estimated densities: bootstrap 90% CI of the adjusted estimator



Correspondingly, the valuation x is obtained by

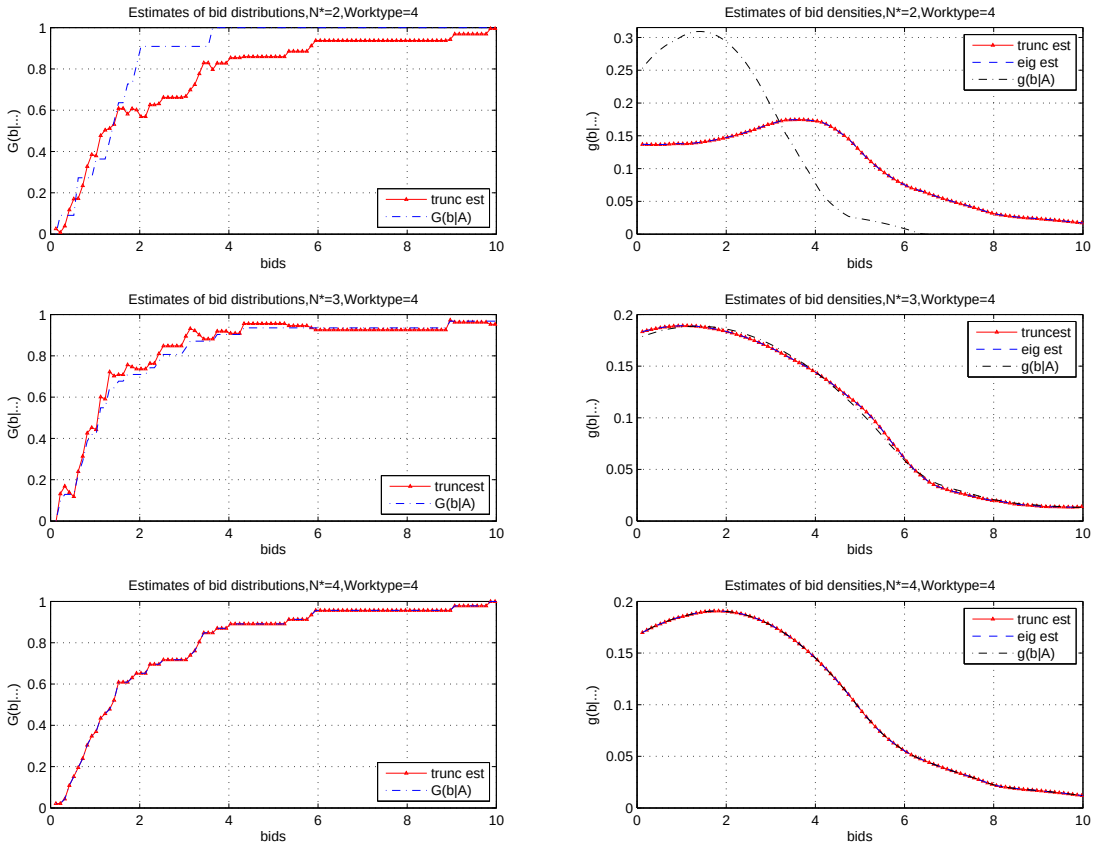
$$\xi\left(b, N^*\right) = b - \frac{1}{N^* - 1} \times \frac{1 - F_{N^*}(r)G(b|N^*)}{F_{N^*}(r)g(b|N^*)}. \tag{4.33}$$

Results: Highway work auctions Figure 4.7 contains the graphs of the estimated densities $g(b|N^*)$ for $N^* = 2, 3, 4$, for the highway work auctions. In each column of this table, we present three estimates of each $g(b|N^*)$: (i) the normalized estimate with the negative portions removed, just following the remedy we mentioned above, labeled "trunc est"; (ii) the un-normalized estimate, which includes the negative values for the density, labeled "Orig est"; and (iii) the naïve estimate, given by $g(b|A)$. In each plot, we also include the 5% and 95% pointwise confidence intervals, calculated using bootstrap resampling.[29]

Figure 4.7 shows that the naïve bid density estimates, using A in place of N^*, overweights small bids, which is reminiscent of the Monte Carlo results. As above, the reason for this seems to be that the number of potential bidders N^* exceeds the observed number of bidders A. In the IPV framework, more competition drives down bids, implying that

[29]The asymptotic variance is derived analytically in the appendix. However, it is tedious to compute in practice, which is why we use the bootstrap to approximate the pointwise variance of the density estimates.

Figure 4.8: Highway work projects, estimated distribution functions and densities



using A to proxy for the unobserved level of competition N^* may overstate the effects of competition. Because in this empirical application we do not know and control the data-generating process, these economically sensible differences between the naïve estimates (using $g(b|A)$) and our estimates (using $g(b|N^*)$) serve as a confirmatory reality check on the assumptions underlying our estimator. In order to observe the performances of these estimators closely by comparisons, we also include estimated empirical CDFs and densities for $N^* = 2, 3, 4$ in Figure 4.8.

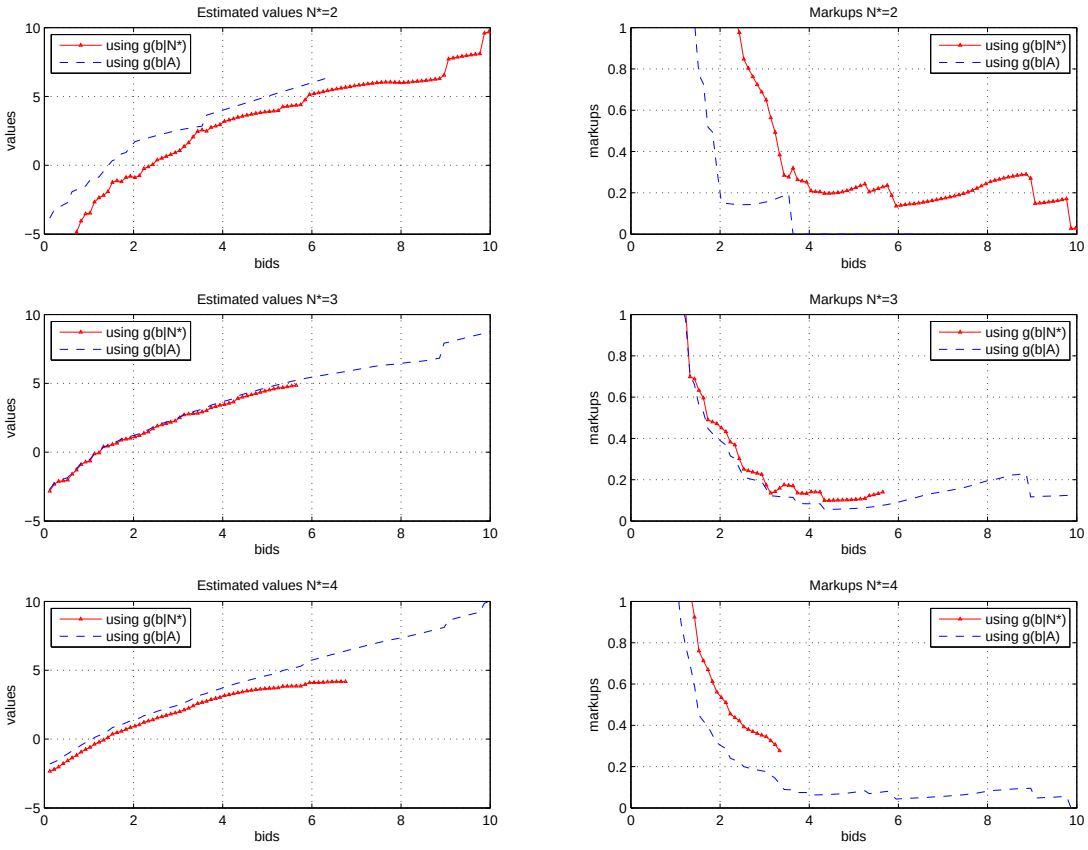
For these estimates, the estimated $G_{A|N^*}$ matrix was

	$N^*=2$	$N^*=3$	$N^*=4$
$A=2$	1.0000	0.1300	0.4091
$A=3$	0	0.8700	0.1041
$A=4$	0	0	0.4868

Furthermore, for the normalized estimates of the bid densities with the negative portions removed, the implied values for $E[b|N^*]$, the average equilibrium bids conditional on N^*, were 3.6726, 3.1567, 3.1776 for, respectively, $N^* = 2, 3, 4$ (in millions of dollars).

The corresponding valuation estimates, obtained by solving Eq. (4.33) pointwise in b

Figure 4.9: Highway work projects, estimated values and markups



using our bid distribution and density estimates, are graphed in Figure 4.9. We present the valuations estimated using our approach, as well as a naïve approach using $g(b|A)$ as the estimate for the bid densities. Note that the valuation estimates become negative within a low range of bids, and then at the upper range of bids, the valuations are decreasing in the bids, which violates a necessary condition of equilibrium bidding. These may be due to unreliability in estimating the bid densities $g(b|A)$ and $g(b|N^*)$ close to the bounds of the observed support of bids.

Comparing the estimates of valuations using $g(b|N^*)$, and those obtained using $g(b|A)$, we see that the valuations using $g(b|N^*)$ are smaller than those using $g(b|A)$, for $N^* = 2, 3, 4$ As in the Monte Carlo results, this implies that the markups $(b - c)/b$ are larger using our estimates of $g(b|N^*)$. The differences in implied markups between these two approaches is economically meaningful, as illustrated in the right-hand-side graphs in Figure (4.9). For example, for $N^* = 4$, at a bid of \$2 million, the corresponding markup using $g(b|A = 4)$ is around 30%, or \$600,000, but using $g(b|N^* = 4)$ is around 55%, or \$1.1 million. This suggests that failing to account for unobservability of N^* can lead the researcher to understate bidders' profit margins.

4.2.6 Extensions

Only Winning Bids are Recorded

In some first-price auction settings, only the winning bid is observed by the researcher. This is particularly likely for the case of descending price, or *Dutch* auctions, which end once a bidder signals his willingness to pay a given price. For instance, Laffont et al. (1995) consider descending auctions for eggplants where only the winning bid is observed, and van den Berg and van der Klaauw (2007) estimate Dutch flower auctions where only a subset of bids close to the winning bid are observed. Within the symmetric IPV setting considered here, Guerre et al. (2000) and Athey and Haile (2002) argue that observing the winning bid is sufficient to identify the distribution of bidder valuations, provided that N^* is known. Our estimation methodology can be applied to this problem even when the researcher does not know N^*, under two scenarios.

First Scenario: Non-Binding Reserve Price In the first scenario, we assume that there is no binding reserve price, but the researcher does not know N^*. (Many Dutch auctions take place too quickly for the researcher to collect data on the number of participants.) Because there is no binding reserve price, the winning bid is the largest out of the N^* bids in an auction. In this case, bidders' valuations can be estimated in a two-step procedure.

In the first step, we estimate $g_{WB}(\cdot|N^*)$, the equilibrium density of winning bids, conditional on N^*, using the methodology above. In the second step, we exploit the fact that in this scenario, the equilibrium CDF of winning bids is related to the equilibrium CDF of the bids by the relation:

$$G_{WB}(b|N^*) = G(b|N^*)^{N^*}.$$

This implies that the equilibrium bid CDF can be estimated as $\hat{G}(b|N^*) = \hat{G}_{WB}(b|N^*)^{1/N^*}$, where $\hat{G}_{WB}(b|N^*)$ denotes the CDF implied by our estimates of $\hat{g}_{WB}(b|N^*)$. Subsequently, upon obtaining an estimate of $\hat{G}(b|N^*)$ and the corresponding density $\hat{g}(b|N^*)$, we can evaluate Eq. (4.9) at each b to obtain the corresponding value.

Second Scenario: Binding Reserve Price, but A Observed In the second scenario, we assume that the reserve price binds, but that A, the number of bidders who are willing to submit a bid above the reserve price, is observed. The reason we require A to be observed is that when reserve prices bind, the winning bid is not equal to $b^{N^*:N^*}$, the highest order statistic out of N^* i.i.d. draws from $g(b|N^*, b > r)$, the equilibrium bid distribution truncated to $[r, +\infty)$. Rather, for a given N^*, it is equal to $b^{A:A}$, the largest out of A i.i.d. draws from $g(b|N^*, b > r)$. Hence, because the density of the winning bid depends on A, even after conditioning on N^*, we must use A as a conditioning covariate in our estimation.

For this scenario, we estimate $g(b|N^*, b > r)$ in two steps. First, treating A as a conditioning covariate, we estimate $g_{WB}(\cdot|A, N^*)$, the conditional density of the winning bids conditional on both the observed A and the unobserved N^*. Second, for a fixed N^*,

we can recover the conditional CDF $G(b|N^*, b > r)$ via

$$\hat{G}(b|N^*, b > r) = \hat{G}_{WB}(b|A, N^*)^{1/A}, \ \forall A.$$

(That is, for each N^*, we can recover an estimate of $G(b|N^*, b > r)$ for each distinct value of A. Since the model implies that these distributions should be identical for all A, we can, in principle, use this as a specification check of the model.)

In both scenarios, we need to find good candidates for the auxiliary variables N and Z. Since typically many Dutch auctions are held in a given session, one possibility for N could be the total number of attendees at the auction hall for a given session, while Z could be an instrument (such as the time of day) which affects bidders' participation for a specific auction during the course of the day.[30]

Endogenous Entry

A second possible extension of our approach is to models of endogenous entry. In Samuelson's (1985) model, N^* potential entrants observe their valuations, and must decide whether or not to pay an entry cost $k > 0$ to bid in the auction. In this model (cf. Li and Zheng (2006), Marmer et al. (2013)), the distribution of the valuations of the bidders who enter the auction, $F_{N^*}(v)$, varies depending on N^*. As Marmer et al. (2013) show, the inverse bidding strategy for this model, analogous to Eq. (4.9), is:

$$\xi(b, N^*) = b + \frac{1 - p(N^*) + p(N^*)G(b|N^*)}{(N^* - 1)p(N^*)g(b|N^*)} \tag{4.34}$$

where $p(N^*)$ denotes the equilibrium entry probability with N^* potential entrants.

We can apply our methodology to identify and estimate the valuation distributions $F_{N^*}(v)$ in this model, even when the number of potential entrants N^* is not observed. Let A denote the number of bidders who enter, which we assume to be observed.[31] First, using our procedure, the equilibrium bid distributions $G(b|N^*)$ and misclassification probabilities $G_{A|N^*}$ can be estimated using A as the proxy for N^* and a second bid in each auction in the role of Z. For recovering the valuations, note that, corresponding to Eq. (4.30), in equilibrium we have

$$A|N^* \ \sim \ \text{Binomial}(N^*, p(N^*)) \tag{4.35}$$

implying that $p(N^*)$ can be recovered for each value of N^* from the misclassification probability matrix $G_{A|N^*}$. Once $p(N^*)$ is known, the valuations can be identified for each b in the support of $G(b|N^*)$ using Eq. (4.34).

[30] This corresponds to the scenario considered in the flower auctions in van den Berg and van der Klaauw (2007).

[31] In this model, a reserve price is irrelevant, because all bidders with valuations below the reserve price will never enter the auction. Hence, we do not need to distinguish between the number of bidders who enter and those who enter and submit a nonzero bid.

4.2.7 Asymptotic Properties of the Two Step Estimator

Proof of uniform consistency of $\widehat{g}(b|N^*)$.

In the first step, we estimate $\widehat{G}_{N|N^*}$ from

$$\widehat{G}_{N|N^*} := \psi\left(\widehat{G}_{Eb,N,Z}\widehat{G}_{N,Z}^{-1}\right), \tag{4.36}$$

where $\psi(\cdot)$ is an analytic function as mentioned in Hu (2008) and

$$\begin{aligned}\widehat{G}_{Eb,N,Z} &= \left[\frac{1}{T}\sum_t \frac{1}{N_j}\sum_{i=1}^{N_j} b_{it}\mathbf{1}(N_t = N_j, Z_t = Z_k)\right]_{j,k},\\ \widehat{G}_{N,Z} &= \left[\frac{1}{T}\sum_t \mathbf{1}(N_t = N_j, Z_t = Z_k)\right]_{j,k}.\end{aligned}$$

We summarize the uniform convergence of $\widehat{G}_{N|N^*}$ as follows:

Lemma 4.2.1 *Suppose that $Var(b|N,Z) < \infty$. Then,*

$$\widehat{G}_{N|N^*} - G_{N|N^*} = O_p\left(T^{-1/2}\right).$$

Proof. It is straightforward to show that $\widehat{G}_{Eb,N,Z} - G_{Eb,N,Z} = O_p\left(T^{-1/2}\right)$ and $\widehat{G}_{N,Z} - G_{N,Z} = O_p\left(T^{-1/2}\right)$. As mentioned in Hu (2008), the function $\psi(\cdot)$ is an analytic function. Therefore, the result holds. ■

In the second step, we have

$$\widehat{g}(b|N^*) = \frac{e_{N^*}^T \widehat{G}_{N|N^*}^{-1}\overrightarrow{\widehat{g}}(b,N)}{e_{N^*}^T \widehat{G}_{N|N^*}^{-1}\overrightarrow{\widehat{g}}(N)},$$

where

$$\widehat{g}(b,N_j) = \frac{1}{Th}\sum_t \frac{1}{N_j}\sum_{i=1}^{N_j} K\left(\frac{b-b_{it}}{h}\right)\mathbf{1}(N_t = N_j).$$

Let $\omega := (b,N)$. Define the norm $\|\cdot\|_\infty$ as

$$\|\widehat{g}(\cdot|N^*) - g(\cdot|N^*)\|_\infty = \sup_b \left|\widehat{g}_{b|N^*}(b|N^*) - g_{b|N^*}(b|N^*)\right|.$$

The uniform convergence of $\widehat{g}(\cdot|N^*)$ is established as follows:

Lemma 4.2.2 *Suppose:*

(2.1) $\omega \in \mathcal{W}$ and $\mathcal{W}$ is a compact set.

(2.2) $g_{b,N}(\cdot,N_j)$ is positive and continuously differentiable to order R with bounded derivatives on an open set containing $\mathcal{W}$.

(2.3) $K(u)$ *is differentiable of order* R*, and the derivatives of order* R *are bounded.* $K(u)$ *is zero outside a bounded set.*$\int_{-\infty}^{\infty} K(u)du = 1$ *, and there is a positive integer* m *such that for all* $j < m$*,* $K^{(j)}(u)$ *is absolutely continuous,* $\int_{-\infty}^{\infty} K(u)u^j du = 0$*, and* $\int_{-\infty}^{\infty} |u|^m |K(u)|du < \infty$.

(2.4) $h = cT^{-\delta}$ *for* $0 < \delta < 1/2$*, and* $c > 0$.

Then, for all j*,*

$$\left\| \widehat{g}(\cdot|N^*) - g(\cdot|N^*) \right\|_\infty = O_p \left\{ \left(\frac{Th}{\ln T} \right)^{-1/2} + h^m \right\}. \tag{4.37}$$

The most important assumption for lemma 4.2.2 is (2.2), which places smoothness restrictions on the joint density $g(b, N)$. Via Eq. (4.13), this distribution is a mixture of conditional distributions $g(b|N^*)$, which possibly have a different support for different N^*. When the supports of $g(b|N^*)$ are known, condition (2.2) only requires the smoothness of $g(b|N^*)$ on its own support $[r, u_{N^*}]$ because the distribution $g(b|N)$ can be estimated piecewise on $[r, u_2], [u_2, u_3], ..., [u_{K-1}, u_K]$. When the supports of $g(b|N^*)$ are unknown, condition (2.2) would require that the density $g(b|N^*)$ for each value of N^* to be smooth at the upper boundary.[32]

Proof. By lemma 4.2.1, it is straightforward to show that

$$\begin{aligned} \widehat{\Pr(N^*)} &= e_{N^*}^T \widehat{G}_{N|N^*}^{-1} \overrightarrow{\widehat{g}}(N) \\ &= e_{N^*}^T G_{N|N^*}^{-1} \overrightarrow{g}(N) + O_p\left(T^{-1/2}\right) \end{aligned}$$

Taking into account the fact that $\overrightarrow{\widehat{g}}(b, N)$ is bounded above, and $\widehat{\Pr(N^*)}$ is of order 1, we conclude that

$$\widehat{g}(b|N^*) = \frac{e_{N^*}^T G_{N|N^*}^{-1} \overrightarrow{\widehat{g}}(b, N)}{e_{N^*}^T G_{N|N^*}^{-1} \overrightarrow{g}(N)} + O_p\left(T^{-1/2}\right),$$

In order to show the consistency of our estimator $\widehat{g}(b|N^*)$, we need the uniform convergence of $\widehat{g}(\cdot, N_j)$. The kernel density estimator has been studied extensively. Following results from lemma 5.4 and the discussion followed in Fan and Yao (2005) , under assumptions of lemma 2, we have for all j[33]

$$\sup_b \left| \widehat{g}_{b,N}(\cdot, N_j) - \mathbb{E}\widehat{g}_{b,N}(\cdot, N_j) \right| = O_p\left(\frac{Th}{\ln T} \right)^{-1/2}. \tag{4.38}$$

According to the discussion on Page 205 in Fan and Yao (2005), assumption (2.3) implies

[32] In ongoing work, we are exploring alternative methods, based on wavelet methods (eg. Hall et al. (1996)), to estimate the joint density $g(b, N)$ when there are unknown points of discontinuity, which can be due to the non-smoothness of the individual densities $g(b|N^*)$ at the upper boundary of their supports.

[33] The results in Fan and Yao (2005) are for $m = 2$ but they also hold for $m > 2$.

that the bias

$$\mathbb{E}\widehat{g}_{b,N}(\cdot, N_j) - g_{b,N}(\cdot, N_j) = O_p(h^m). \tag{4.39}$$

Consider that

$$\left|\widehat{g}_{b,N}(\cdot, N_j) - g_{b,N}(\cdot, N_j)\right| \leq \left|\widehat{g}_{b,N}(\cdot, N_j) - \mathbb{E}\widehat{g}_{b,N}(\cdot, N_j)\right| + \left|\mathbb{E}\widehat{g}_{b,N}(\cdot, N_j) - g_{b,N}(\cdot, N_j)\right|.$$

From (4.38) and (4.39), we immediately conclude that

$$\sup_b \left|\widehat{g}_{b,N}(\cdot, N_j) - g_{b,N}(\cdot, N_j)\right| = O_p\left\{\left(\frac{Th}{\ln T}\right)^{-1/2} + h^m\right\}.$$

The uniform convergence of $\widehat{g}_{b|N^*}$ then follows. ■

Remark: Another technical issue pointed out in Guerre et al. (2000) is that the density $g(b|N^*)$ may not be bounded at the lower bound of its support, which is the reserve price r. They suggest using the transformed bids $b_\dagger \equiv \sqrt{b-r}$. Our identification and estimation procedures remain the same if b replaced by $b_\dagger$, where an estimate of the reserve price r could be the lowest observed bid in the dataset (given our assumption that the reserve price is fixed in the dataset). ■

Proof of asymptotic normality of $\widehat{g}(b|N^*)$ In this section, we show the asymptotic normality of $\widehat{g}(b|N^*)$ for a given value of b. Define $\gamma_0(b) = \{g_{b,N}(b)\}$, a column vector containing all the elements in the matrix $g(b, N)$. Similarly, we define $\widehat{\gamma}(b) = \{\widehat{g}_{b,N}(b)\}$. The proof of Lemma 4.2.2 suggests that

$$\widehat{g}(b|N^*) = \varphi(\widehat{\gamma}(b)) + O_p\left(T^{-1/2}\right)$$

where

$$\varphi(\widehat{\gamma}(b)) \equiv \frac{e_{N^*}^T G_{N|N^*}^{-1} \overrightarrow{\widehat{g}}(b, N)}{e_{N^*}^T G_{N|N^*}^{-1} \overrightarrow{g}(N)}.$$

Notice that the function $\varphi(\cdot)$ is linear in each entry of the vector $\widehat{\gamma}(b)$. Therefore, we have

$$\widehat{g}(b|N^*) - g(b|N^*) = \left(\frac{d\varphi}{d\gamma}\right)^T (\widehat{\gamma}(b) - \gamma_0(b)) + o_p\left(1/\sqrt{Th}\right),$$

where $\frac{d\varphi}{d\gamma}$ is nonstochastic because it is a function of $G_{N|N^*}$ and $\overrightarrow{g}(N)$ only. The asymptotic distribution of $\widehat{g}(b|N^*)$ then follows that of $\widehat{\gamma}(b)$. We summarize the results as follows:

Lemma 4.2.3 *Suppose that assumptions in Lemma 4.2.2 hold with $R = 2$ and that*

1. *there exists some δ such that $\int |K(u)|^{2+\delta} du < \infty$,*
2. *$(Th)^{1/2} h^2 \to 0$, as $T \to \infty$.*

Then, for a given b and N^*,

$$(Th)^{1/2} [\widehat{g}(b|N^*) - g(b|N^*)] \xrightarrow{d} N(0, \Omega),$$

where

$$\begin{aligned}\Omega &= \left(\frac{d\varphi}{d\gamma}\right)^T V\left(\widehat{\gamma}\right)\left(\frac{d\varphi}{d\gamma}\right), \\ V\left(\widehat{\gamma}\right) &= \lim_{T\to\infty}(Th)\, E\left[\left(\widehat{\gamma}-E\left(\widehat{\gamma}\right)\right)\left(\widehat{\gamma}-E\left(\widehat{\gamma}\right)\right)^T\right].\end{aligned}$$

Proof. As discussed above, the asymptotic distribution of $\widehat{g}(b|N^*)$ is derived from that of $\widehat{\gamma}(b)$. In order to prove that the asymptotic distribution of the vector $\widehat{\gamma}(b)$ is multivariate normal $N\left(0, V\left(\widehat{\gamma}\right)\right)$, we show that the scalar $\lambda^T\widehat{\gamma}(b)$ for any vector λ has a normal distribution $N\left(0, \lambda^T V\left(\widehat{\gamma}\right)\lambda\right)$. For a given value of b, it is easy to follow the proof of Theorems 2.9 and 2.10 in Pagan and Ullah (1999) to show that

$$(Th)^{1/2}\left[\lambda^T\widehat{\gamma}(b)-\lambda^T\gamma_0(b)\right] \xrightarrow{d} N\left(0, Var\left(\lambda^T\widehat{\gamma}(b)\right)\right),$$

where $Var\left(\lambda^T\widehat{\gamma}(b)\right) = \lambda^T V\left(\widehat{\gamma}(b)\right)\lambda$ is the variance of the scalar $\lambda^T\widehat{\gamma}(b)$. The asymptotic distribution of $\widehat{g}(b|N^*)$ then follows. ■

4.3 Beliefs in Learning Models

How economic agents learn from past experience has been an important issue in both empirical industrial organization and labor economics. (See Ching et al. (2013) and Ching et al. (2017) for a review.)The key difficulty in the estimation of learning models is that beliefs are time-varying and unobserved in the data. Hu et al. (2013b) use bandit experiments to non-parametrically estimate the learning rule using auxiliary measurements of beliefs. In each period, an economic agent is asked to choose between two slot machines, which have different winning probabilities. Based on her own belief on which slot machine has a higher winning probability, the agent makes her choice of slot machine and receives rewards according to its winning probability. Although she does not know which slot machine has a higher winning probability, the agent is informed that the winning probabilities may switch between the two slot machines.

In addition to choices Y_t and rewards R_t, researchers also observe a proxy Z_t for the agent's belief X_t^*. Recorded by an eye-tracker machine, the proxy describes how much more time the agent looks at one slot machine than at the other. Under a first-order Markovian assumption, the learning rule is described by the distribution of the next period's belief conditional on previous belief, choice, and reward, i.e., $\Pr\left(X_{t+1}^*|X_t^*, Y_t, R_t\right)$. They assume that the choice only depends the belief and that the proxy Z_t is also independent of other variables conditional on the current belief X_t^*. The former assumption is motivated by a fully-rational Bayesian belief-updating rule, while the latter is a local independence assumption widely-used in the measurement error literature. These assumptions imply a 2.1-measurement model with

$$Z_t \perp Y_t \perp Z_{t-1}|X_t^*. \tag{4.40}$$

Therefore, the proxy rule $\Pr\left(Z_t|X_t^*\right)$ is non-parametrically identified. Under the local in-

dependence assumption, one can identify distribution functions containing the latent belief X_t^* from the corresponding distribution functions containing the observed proxy Z_t. That means the learning rule $\Pr\left(X_{t+1}^*|X_t^*, Y_t, R_t\right)$ can be identified from the observed distribution $\Pr\left(Z_{t+1}, Y_t, R_t, Z_t\right)$ through

$$\begin{aligned} & \Pr\left(Z_{t+1}, Y_t, R_t, Z_t\right) \\ = & \sum_{X_{t+1}^*} \sum_{X_t^*} \Pr\left(Z_{t+1}|X_{t+1}^*\right) \Pr\left(Z_t|X_t^*\right) \Pr\left(X_{t+1}^*, X_t^*, Y_t, R_t\right). \end{aligned} \tag{4.41}$$

The nonparametric learning rule they found implies that agents are more reluctant to "update down" following unsuccessful choices, than "update up" following successful choices. That leads to the sub-optimality of this learning rule in terms of profits. We provide details in Hu et al. (2013b), which investigate how individuals learn from past experience in dynamic choice environments.

A growing literature has documented, using both experimental and field data, that the benchmark fully-rational Bayesian learning model appears deficient at characterizing actual decision-making in real-world settings. Other papers have demonstrated that observed choices in strategic settings with asymmetric information are typically not consistent with subjects' having Bayesian (equilibrium) beliefs regarding the private information of their rivals. Recently, non-Bayesian *reinforcement learning* (Sutton and Barto (1998)) models have also been used to explain some observed anomalies in savings and investment behavior (eg. Choi et al. (2009), Odean et al. (2004)).

Given the lack of consensus in the literature about what the actual learning rules used by agents in real-world decision environments are, there is a need for these rules to be estimated in a manner flexible enough to accommodate alternative models of learning. In this paper, we propose a new approach for assessing agents' belief dynamics. In an experimental setting, we utilize data on subjects' *eye-movements* during the experiment to aid our inference regarding the learning (or belief-updating) rules used by subjects in their decision-making process. Previous studies have established a connection between subjects' eye movements and gaze durations and their valuations in choice experiments. We exploit this connection and use gaze durations to *pin down* subjects' evolving beliefs in a dynamic choice setting.

This paper is the first to use such "neuroeconomic" data in estimating behavioral decision-making models. Taking advantage of recent developments in the econometrics of dynamic measurement error models, we use the observed choice and eye- tracking data to estimate subjects' decision rules and learning rules, without imposing a priori functional forms on these functions. Estimating the learning rules in such a model-free manner allows us to assess the optimality of subjects' choices in learning experiments in a manner quite distinct from that taken in the existing literature.

The main results are as follows. First, our estimated learning rules do not correspond to any one of the existing learning models. Rather, we find that beliefs are reward-asymmetric, in that subjects are more reluctant to "update down" following unsuccessful (low-reward) choices, than "update up" following successful (high-reward) choices. Such asymmetries

are novel relative to existing learning models (such as reinforcement or Bayesian learning); moreover, from a payoff perspective, they are suboptimal relative to the fully-rational Bayesian benchmark.

Correspondingly, we find that, using the estimated learning rules, subjects' payoffs are, at the median, \$4 (or about two cents per choice) lower than under the Bayesian benchmark; this difference represents about 25% of typical experimental earnings (not including the fixed show-up fee). However, subjects' payoffs under the estimated choice and learning rules are comparable to the profits from alternative non-Bayesian learning models, including reinforcement learning.

4.3.1 Two-Armed Bandit "Reversal Learning" Experiment

Our experiments are adapted from the "reversal learning" experiment used in Hampton et al. (2006). In the experiments, subjects make repeated choices between two actions (which we call interchangeably "arms" or "slot machines" in what follows): in trial t, the subject chooses $Y_t \in \{1(=\text{"green"}), 2(=\text{"blue"})\}$. The rewards generated by these two arms are changing across trials, as described by the state variable $S_t \in \{1, 2\}$, which is never observed by subjects. When $S_t = 1$, then green (blue) is the "good" ("bad") state, whereas if $S_t = 2$, then blue (green) is the "good" ("bad") state.

The rewards R_t that the subject receives in trial t depends on the action taken, as well as (stochastically) on the current state: the reward process is

$$R_t = \begin{cases} \pm\$0.50 \text{ with prob. } 50\% \pm 20\% & \text{if good arm chosen} \\ \pm\$0.50 \text{ with prob. } 50\% \mp 10\% & \text{if bad arm chosen.} \end{cases} \tag{4.42}$$

For convenience, we use the notation $R_t = 1$ to denote the negative reward (-\$0.50), and $R_t = 2$ to denote the positive reward (\$0.50).

The state evolves according to an exogenous binary Markov process. At the beginning of each block, the initial state $S_1 \in \{1, 2\}$ is chosen with probability 0.5, randomly across all subjects and all blocks. Subsequently, the state evolves with transition probabilities[34]

$P(S_{t+1}\|S_t)$	$S_t = 1$	$S_t = 2$
$S_{t+1} = 1$	0.85	0.15
$S_{t+1} = 2$	0.15	0.85

. (4.43)

Because S_t is not observed by subjects, and is serially-correlated over time, subject have an opportunity to learn and update their beliefs about the current state on the basis of past rewards. Moreover, because S_t changes randomly over time, so that the identity of the good arm varies across trials, this is called a "probabilistic reversal learning" experiment.

[34]This aspect of our model differs from Hampton et al. (2006), who make the non-Markovian assumption that the state S_t changes with probability 25% after a subject has chosen the good arm four successive times. Estimating such non-Markovian models would require alternative identification arguments than the one considered in this paper.

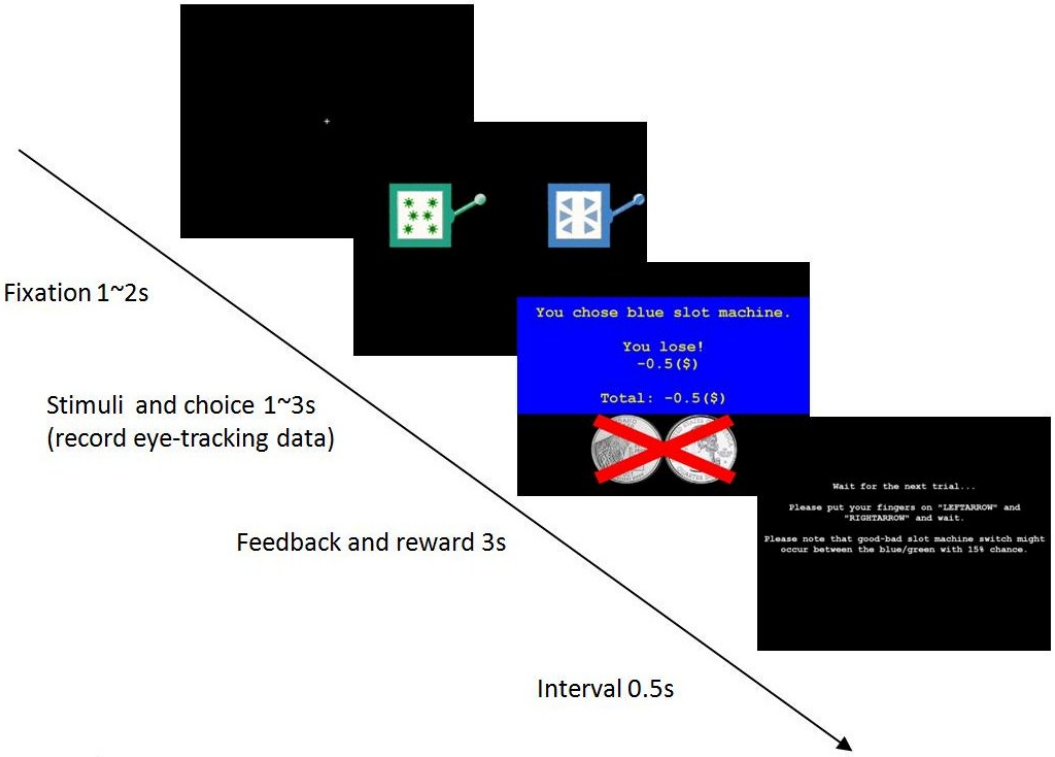


Figure 4.10: Timeline of a trial

After subjects fix their gaze on the cross (top screen), two slot machines are presented (second screen). Subjects' eye-movements are recorded by the eye-tracking machine here. Subjects choose by pressing the left (right) arrow key to indicate a choice of the left (right) slot machine. After choosing (third screen), a positive reward (depicted by two quarters) or negative reward (two quarters covered by a red X) is delivered, along with feedback about the subject's choice highlighted against a background color corresponding to the choice. In the bottom screen, a subject is transitioned to the next trial, and reminded that a slot machine may switch from "good" to "bad" (and vice versa) with probability 15%.

Table 4.2: Summary statistics for experimental data

	1(green)	2(blue)
Y: subjects' choices	2108	2092

	2 ($0.50)	1 (-$0.50)
R: rewards	2398	1802

	mean	median	upper 5%	lower 5%
$\tilde{Z}$: eye movement measure[a]	-0.0309	0	1.3987	-1.4091
RT: reaction time (10^{-2} secs)	88.22	59.3	212.2	36.8

[a]Defined in Eq. (4.44)

Experimental Data: Preliminary Analysis

The experiments were run over several weeks in November-December 2009. We used 21 subjects, recruited from the Caltech Social Science Experimental Laboratory (SSEL) subject pool consisting of undergraduate/graduate students, post-doctoral students, and community members,[35] each playing for 200 rounds (broken up into 8 blocks of 25 trials). Most of the subjects completed the experiment within 40 minutes, including instruction and practice sessions. Subjects were paid a fixed show-up fee ($20), in addition to the amount won during the experiment, which was $14.20 on average.[36]

Subjects were informed of the reward structure for good and bad slot machines, and the Markov transition probabilities for state transitions (reversals), but were not informed which state was occurring in each trial. Figure 4.10 contains the time line and some screenshots from the experiment. In addition, while performing the experiment, the subjects were attached to an eye-tracker machine, which recorded their eye movements. From this, we constructed the auxiliary variable $\tilde{Z}_t$, which measures the fraction of the reaction time (the time between the onset of a new round after fixation, and the subject's choice in that round) spent gazing at the picture of the "blue" slot machine on the computer screen.[37]

For each subject, and each round t, we observe the data (Y_t, S_t, R_t, Z_t). Table 4.2 presents some summary statistics of the data. The top panel shows that, across all subjects and all trials, "green" (2108 choices) and "blue" (2092 choices) are chosen in almost-equal

[35]Community members consisted of spouses of students at either Caltech or Pasadena City College (a two-year junior college). While the results reported below were obtained by pooling the data across all subjects, we also estimated the model separately for the subsamples of Caltech students, vs. community members. There were few noticeable differences in the results across these classes of subjects.

[36]For comparison, purely random choices would have earned $10 on average.

[37]Across trials, the location of the "blue" and "green" slot machines were randomized, so that the same color is not always located on the same side of the computer screen. This controls for any "right side bias" which may be present (see discussion further below).

proportions. Moreover, from the second panel, we see that subjects obtain the high reward with frequency of roughly 57% ($\approx 2398/(2398+1802)$). This is slightly higher than, but significantly different from, 55%, which is the frequency which would obtain if the subjects were choosing completely randomly.[38] Hence, subjects appear to be "trying", which motivates our analysis of their learning rules. On the other hand, simulation of the fully-rational Bayesian decision rules (discussed above) show that the success rate from using the fully-rational decision rule is only 58.4%, which is just slightly higher than the in-sample success rate found in the experiments. It appears, then, that in the reversal learning setting, the success rate intrinsically varies quite narrowly between 55% and 58.4%.

Table 4.3 contains the empirical frequencies of choices in period t, conditional on choices and rewards from the previous period $(Y_t|Y_{t-1}, R_{t-1})$. This can be interpreted as a "reduced-form" decision rule for the subjects. The top row in that table contains the empirical frequencies, across all subjects, that the green arm is chosen, conditional on the different values of $(Y_t|Y_{t-1}, R_{t-1})$. Looking at the second (fourth) entry in this row, we see that after a successful choice of green (blue), a subject replays this strategy with probability 0.86 (0.88=1-0.12). Thus, on average, subjects appear to replay successful strategies, corresponding to a "win-stay" rule-of-thumb.

However subjects appear reluctant to give up *unsuccessful* strategies. The probability of replaying a strategy after an unsuccessful choice of the same strategy is around 50% for both the blue and green choices (ie. the first and third entries in this row). Thus, subjects tend to randomize after unsuccessful strategies. As far as we are aware, such an "asymmetric" choice rule is new in the literature; moreover, as we will see below, this is echoed in the "asymmetric" belief-updating rule which we estimate.

In the remainder of Table 4.3, we also present the same empirical frequencies, calculated for each subject individually. There is some degree of heterogeneity in subjects' strategies. Looking at columns 2 and 4 or the table, we see that, for the most part, subjects pursue a "win-stay" strategy: the probabilities in the second column are mainly $>>$ 50%, and those in the fourth column are most $<<$ 50%. However, looking at columns 1 and 3, we see that there is significant heterogeneity in subjects' choices following a low reward. In these cases, randomization (which we classify as a choice probability between 40-60%) is the modal strategy among subjects; strikingly, however, a number of subjects continue replaying an unsuccessful strategy: for examples, subjects 3,8, and 11 continue to choose "green" with probabilities of 79%, 89% and 79% even after a previous choice of green yielded a negative reward.[39]

We define $\tilde{Z}_{it}$, our raw eye-movement measure, as the difference in the gaze duration directed at the blue and green slot machines, normalized by the total reaction time:

$$\tilde{Z}_t = (Z_{b,t} - Z_{g,t})/RT_t; \tag{4.44}$$

[38] This is the marginal probability of a good reward, which equals $0.5(0.7+0.4)$ from Eq. (4.42). The t-statistic for the null that subjects are choosing randomly equals 169.67, so that hypothesis is strongly rejected.

[39] In the reversal learning model, however, such a strategy is not obviously irrational; because the identity of the good arm changes exogenously across periods, an arm that was bad last period (ie. yielding a low reward) may indeed be good in the next period.

Table 4.3: "Reduced form" decision rules: $P(Y_t = 1(green)|Y_{t-1}, R_{t-1})$
Choice probabilities conditional on past choice Y_{t-1} and reward R_{t-1}

(Y_{t-1}, R_{t-1})=	(1,1)	(1,2)	(2,1)	(2,2)
Across All Subjects:	0.5075	0.8652	0.5089	0.1189
	(0.0169)	(0.0094)	(0.1169)	(0.0090)
For each individual subject:				
Subject #1:	0.1799	0.5192	0.8128	0.364
	(0.0655)	(0.0684)	(0.0595)	(0.0603)
Subject #2:	0.1051	0.9820	0.9449	0
	(0.0498)	(0.0171)	(0.0381)	(0)
Subject #3:	0.7938	0.9859	0.3340	0
	(0.0591)	(0.0136)	(0.0871)	(0)
Subject #4:	0.3244	0.8796	0.6492	0.0610
	(0.0704)	(0.0514)	(0.0726)	(0.0283)
Subject #5:	0.0419	0.8796	0.6492	0.0610
	(0.0292)	(0.0236)	(0.0325)	(0.0461)
Subject #6:	0.2570	0.7498	0.8159	0.2021
	(0.0652)	(0.0592)	(0.0602)	(0.0532)
Subject #7:	0.5792	0.9242	0.4647	0.0796
	(0.0751)	(0.0371)	(0.0731)	(0.0379)
Subject #8:	0.8931	0.9803	0.1013	0.0165
	(0.0496)	(0.0186)	(0.0482)	(0.0163)
Subject #9:	0.6377	1.0000	0.2741	0
	(0.0831)	(0)	(0.0655)	(0)
Subject #10:	0.1986	0.9344	0.8037	0
	(0.0622)	(0.0352)	(0.0587)	(0)
Subject #11:	0.7859	1.0000	0.4306	0
	(0.0575)	(0)	(0.0870)	(0)
Subject #12:	0.5883	0.9262	0.3741	0.0131
	(0.0841)	(0.0406)	(0.0733)	(0.0129)
Subject #13:	0.6741	0.8907	0.1962	0.2085
	(0.0705)	(0.0462)	(0.0581)	(0.0539)
Subject #14:	0.4730	0.6147	0.5363	0.3842
	(0.0831)	(0.0653)	(0.0735)	(0.0664)
Subject #15:	0.6759	0.9789	0.3351	0
	(0.0761)	(0.0206)	(0.0714)	(0)
Subject #16:	0.4595	0.9135	0.5443	0.1953
	(0.0715)	(0.0316)	(0.0742)	(0.0666)
Subject #17:	0.6358	0.5202	0.5322	0.4644
	(0.0660)	(0.0706)	(0.0780)	(0.0748)
Subject #18:	0.6333	1.0000	0.2901	0
	(0.0834)	(0)	(0.0734)	(0)
Subject #19:	0.6144	0.8197	0.5808	0.2013
	(0.0702)	(0.0444)	(0.0806)	(0.0625)
Subject #20:	0.3699	0.5741	0.3699	0.3554
	(0.0858)	(0.0707)	(0.0665)	(0.0621)
Subject #21:	0.6990	0.9602	0.2934	0.0177
	(0.0658)	(0.0274)	(0.0693)	(0.0171)

Note: standard errors (in parentheses) computed using 1000 block-bootstrap resamples

that is, for trial t, $Z_{b(g),t}$ is the gaze duration at the blue (green) slot machine, and RT_t is the reaction time, ie. the time between the onset of the trial after fixation, and the subject's choice.[40] Thus, $\tilde{Z}_t$ measures how much longer a subject looks at the blue slot machine than the green one during the t-th trial, with a larger (smaller) value of $\tilde{Z}_t$ implying longer gazes directed at the blue (green) slot machine. Summary statistics on this measure are given in the bottom panel of Table 4.2. There, we see that the average reaction time is 0.88 seconds, and that the median value of $\tilde{Z}_t$ is zero, implying an equal amount of time directed to each of the two slot machines.

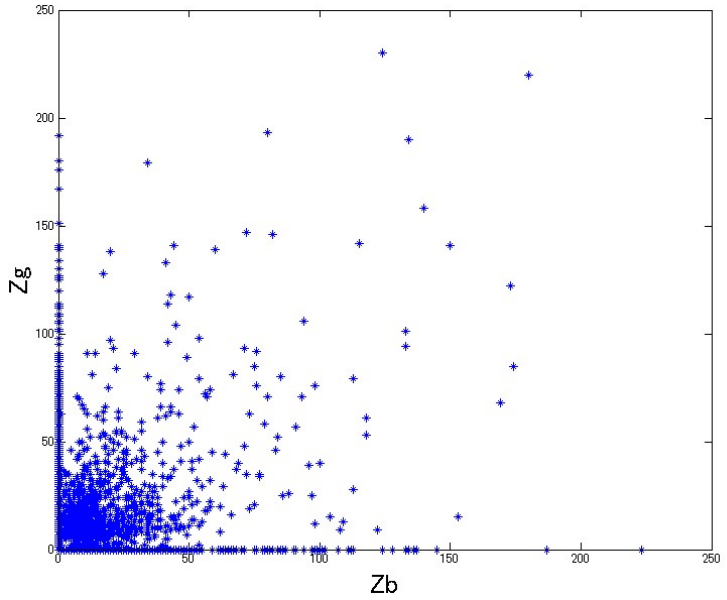


Figure 4.11: Scatter plot of Z_b (gaze at blue) and Z_g (gaze at green)
Both Z_b and Z_g are reported in 2×10^{-2} seconds.

Figure 4.11 contains the scatter plot of $Z_{b,t}$ versus $Z_{g,t}$. In our empirical work, we will discretize the eye-movement measure $\tilde{Z}_t$; to avoid confusion, in the following we use $\tilde{Z}_t$ to denote the undiscretized eye-movement measure, and Z_t the discretized measure, which we describe below.

4.3.2 Empirical Model

In this section, we introduce a model of dynamic decision-making in the two-armed bandit experiment described above, and also discuss the identification and estimation of this model. Because the gamut of learning models is very large, we start by discussing a benchmark model – the fully-rational Bayesian learning model. After considering that model, we describe the empirical model which we take to the data, which allows for deviations from the fully-rational benchmark. Importantly, in our empirical work, we will not consider the whole gamut of learning models, but restrict attention to models which are "close" to

[40]Furthermore, in order to control for subject-specific heterogeneity, we normalize $\tilde{Z}_t$ across subjects by dividing by the subject-specific standard deviation of $\tilde{Z}_t$, across all rounds for each subject.

fully-rational in that the structure of the learning and decision rules are the same as in the benchmark fully-rational model; however, the rules themselves are allowed to be different.

Benchmark: Fully-rational Decision-Making in Reversal Learning Model

As in the experiments, we consider a finite (25 period) dynamic optimization problem, in which each subject aims to choose a sequence of actions to maximize expected rewards $\mathbb{E}\left[\sum_{t=1}^{25} R_t\right]$. (The details of this model are given in Appendix A.)

Let B_t^* denote the probability (given by Bayes' Rule) that a subject places on "blue" being the good arm in period t, conditional on the whole experimental history up to then. We evaluate the fully-rational decision rules – the mapping from period t beliefs B_t^* to a period t choice – in this dynamic Bayesian learning model by computer simulation. Importantly, we accommodate nonstationarity in the problem, in that our simulations allow the decision rules to differ arbitrarily across periods. This permits the relationship between subjects' choices and their beliefs B_t^* to vary across periods, depending perhaps on the periods remaining in the experiment, or to allow for history dependence in either choices or the belief-updating rule. An important maintained assumption in this paper is that subjects' decision rules are solely a function of the current state probabilities B_t^*, so that by allowing the decision rules to vary across periods in these simulations, we can assess the restrictiveness of such an assumption.

We plot the optimal decision rules for this model . Two features are apparent. First, we see that the decision rules are identical across all the periods, indicating that they are *stationary*. Second, the fully-rational decision rule takes a simple form: in each period, it prescribes that subjects choose the blue arm whenever the current belief B_t^* that the blue arm is "good" exceeds 50%. This is a *myopic* decision rule.

These optimal decision rules for the reversal learning model differ in important ways from optimal decision-making in the standard multi-armed bandit (MAB) problem (cf. Gittins and Jones (1974), Banks and Sundarum (1992)), in which the states of the bandits are fixed over all periods and the bandits are "independent" in that a reward from one bandit is uninformative about the state of another bandit. The Bayesian decision rule in the standard MAB model features exploration (or "experimentation"), which recommends sacrificing current rewards to achieve longer-term payoffs; this makes simple myopic decision-making (choosing the bandit which currently has the higher expected reward) suboptimal.[41] In our reversal learning setting, however, the states of the bandits are negatively related, so that positive information about one arm implies negative information about the other.

The Empirical Model

Having described fully-rational behavioral benchmark in our experimental setup, we now describe the empirical model which we fit to the experimental data. As we remarked earlier, the assumptions of this empirical model will be motivated by the nature of decision-making

[41] See Miller (1984), Erdem and Keane (1996), Ackerberg (2003), Crawford and Shum (2005), Chan and Hamilton (2006), and Marcoul and Weninger (2008) for empirical studies of learning and experimentation in a dynamic choice context.

in the benchmark model. In that sense, we do not allow subjects to make arbitrarily irrational decisions, but rather to use decision and belief-updating rules which are "close" to fully-rational.

We introduce the variable X_t^*, which denotes the subject's round t beliefs about the current state S_t; obviously, subjects know their beliefs X_t^*, but these are unobserved by the researcher.[42] In what follows, we assume that both X^* and Z are discrete, and take support on K distinct values which, without loss of generality, we denote $\{1, 2, \ldots, K\}$. We make the following assumptions regarding the subjects' learning and decision rules:

Assumption 4.3.1 *Subjects' choice probabilities $P(Y_t|X_t^*)$ only depend on current beliefs. Moreover, the choice probabilities $P(Y_t = y|X_t^*)$ varies across different values of X_t^* (ie. beliefs affect actions).*

Because we interpret the unobserved variables X_t^* here as a reflection of subjects' *current* beliefs regarding which arm is currently the "good" one, the choice probability $P(Y_t|X_t^*)$ can be interpreted as that which arises from a "myopic" choice rule. As we remarked before, in Section 1.1, such an interpretation is justified by the simulation of the fully-rational decision rules under the reversal learning setting, which showed that these rules are myopic and depend only on current beliefs.

This assumption embodies the core of our strategy for estimating subjects' beliefs; it posits important exclusion restrictions that, conditional on beliefs X_t^*, the observed action Y_t is independent of everything else, including the eye movement Z_t as well as past choices Y_{t-1}. Table 4.3 showed that choices are serially correlated across periods; assumption 1 implies that this serially correlation is due entirely to the unobserved beliefs X_t^* — thus, beliefs (which are unobserved to the researcher) are the reason for serial correlation in choices observed in Table 4.3.

Assumption 4.3.2 *The law of motion for X_t^*, which describes how subjects' beliefs change over time given the past actions and rewards, is called the **learning rule**. This is a controlled first-order Markov process, with transition probabilities $P(X_t^*|X_{t-1}^*, R_{t-1}, Y_{t-1})$.*

This assumption is motivated by the structure of the fully-rational Bayesian belief-updating rule (cf. Eq. (4.50) in Appendix A), in which the period t beliefs depend only on the past beliefs, actions, and rewards in period $t-1$. However, we allow the exact form of the learning rule to deviate from the exact Bayes formula.

Assumption 4.3.3 *The eye movement measure Z_t is a noisy measure of beliefs X_t^*:*
(i) Eye movements are serially uncorrelated conditional on beliefs: $P(Z_t|X_t^*, Y_t, Z_{t-1}) = P(Z_t|X_t^*)$.
(ii) For all t, the $K \times K$ matrix $\mathbf{G}_{Z_t|Z_{t-1}}$, with $(i,j)-th$ entry equal to $P(Z_t = i|Z_{t-1} = j)$, is invertible.
(iii) $E[Z_t|X_t^*]$ is increasing in X_t^*.

[42] X_t^* corresponds to the prior beliefs p_t from the previous section except that, further below, we will discretize X_t^* and assume that it is integer-valued. Therefore, to prevent any confusion, we will use distinct notation p_t, X_t^* to denote, respectively, the beliefs in the theoretical vs. the empirical model.

Assumption 4.3.3 involves an important exclusion restriction that, conditional on X_t^*, the eye movement Z_t in period t is independent of Z_{t-1}. This assumption is reasonable because, in the experimental setup, we require subjects to "fix" their gaze in the middle of the computer screen at the beginning of each period. This should remove any inherent serial correlation in eye movements which is not related to the learning task.[43]

The invertibility assumption 4.3.3(i) is made on the observed matrix $\mathbf{G}_{Z_t|Z_{t-1}}$ with elements equal to the conditional distribution of $Z_t|Z_{t-1}$; hence it is testable. Assumption 4.3.3(ii) "normalizes" the beliefs X_t^* in the sense that, because large values of Z_t imply that the subject gazed longer at blue, the monotonicity assumption implies that larger values of X_t^* denote more "positive" beliefs that the current state is blue.

Assumption 4.3.4 *The conditional probability distributions describing subjects' choices ($P(Y_t|X_t^*)$), learning rules ($P(X_t^*|X_{t-1}^*, R_{t-1}, Y_{t-1})$), and eye movements ($P(Z_t|X_t^*)$) are the same for all subjects and trials t.*

This stationarity and homogeneity assumption justifies pooling the data across all subjects and trials for estimating the model. Stationarity is motivated by the structure of optimal decision-making discussed above, where both the Bayesian belief-updating rule (Eq. (4.50) in Appendix A) and optimal choice rules are indeed stationary.[44]

Estimation and identification

In the model described in previously, the unknown functions we want to estimate are:
(i) $P(Y_t|X_t^*)$, the *choice probabilities*;
(ii) the *learning rule* $P(X_t^*|X_{t-1}^*, Y_{t-1}, R_{t-1})$; and
(iii) the *eye movement probabilities* $P(Z_t|X_t^*)$, the mapping between the auxiliary measure Z_t and the unobserved beliefs X_t^*.

Despite its simplicity, this model is not straightforward to estimate, because these unknown functions depend on the latent beliefs X_t^*, which are not only unobserved but changing over time.[45] Next, we show formally how the availability of the eye movements Z_t allows us to identify and estimate these unknown functions: essentially, given our assumptions, the eye movements play the role of "noisy measurements" of the underlying latent belief process.[46] Our estimator for the learning rules is very simple, and involves only elementary matrix calculations using matrices which can be formed from the observed data.

For simplicity, we will use the shorthand notation $P(\cdots)$ to denote generically a probability distribution. The identification argument (and , subsequently, estimation procedure)

[43] At the same time, we have also estimated models in which we allow Z_t and Z_{t-1} to be correlated, even conditional on X_t^*. The results, which can be obtained from the authors upon request, indicate that the results are quite similar, for different values of Z_{t-1}, which imply that Assumption 3 is quite reasonable.

[44] The homogeneity assumption could be avoided at the (large) cost of gathering enough data per subject, such that the model could be estimated for each subject individually. Given the eye fatigue facing subjects who are attached to an eye tracker, running so many trials per subject is not feasible.

[45] Specifically, this model is a nonlinear "hidden state Markov" model, which are typically quite challenging to estimate (cf. Ghahramani (2001) and Arcidiacono and Miller (2011)).

[46] We apply recent econometric tools developed for the estimation of nonclassical measurement error models and dynamic discrete-choice models (Hu (2008), Hu and Shum (2012)).

takes two-steps. In the first step, the goal is to recover the choice and eye movement probability functions – that is, the probabilities $P(Y_t|X_t^*)$ (resp. $P(Z_t|X_t^*)$) of a given choice (resp. of given eye gaze duration) conditional on the latent beliefs. In the second step, we recover the learning rules. We describe both steps in turn.

First step. The joint probability distribution $P(Z_t, Y_t|Z_{t-1})$ can be factorized as follows:

$$\begin{aligned} P(Z_t, Y_t|Z_{t-1}) &= \sum_{X_t^*} P(Z_t|Y_t, X_t^*, Z_{t-1}) P(Y_t|X_t^*, Z_{t-1}) P(X_t^*|Z_{t-1}) \\ &= \sum_{X_t^*} P(Z_t|X_t^*) P(Y_t|X_t^*) P(X_t^*|Z_{t-1}) \end{aligned}$$

where the last equality applies assumptions 1 and 3. For any fixed $Y_t = y$, then, we can write the above in matrix notation as:

$$\mathbf{A}_{y,Z_t|Z_{t-1}} = \mathbf{B}_{Z_t|X_t^*} \mathbf{D}_{y|X_t^*} \mathbf{C}_{X_t^*|Z_{t-1}}$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are all $K \times K$ matrices, defined as:

$$\begin{aligned} \mathbf{A}_{y,Z_t|Z_{t-1}} &= \left[P_{Y_t,Z_t|Z_{t-1}}(y, i|j) \right]_{i,j} \\ \mathbf{B}_{Z_t|X_t^*} &= \left[P_{Z_t|X_t^*}(i|k) \right]_{i,k} \\ \mathbf{C}_{X_t^*|Z_{t-1}} &= \left[P_{X_t^*|Z_{t-1}}(k|j) \right]_{k,j} \\ \mathbf{D}_{y|X_t^*} &= \begin{bmatrix} P_{Y_t|X_t^*}(y|1) & 0 & 0 \\ 0 & P_{Y_t|X_t^*}(y|2) & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & P_{Y_t|X_t^*}(y|K) \end{bmatrix} \end{aligned} \tag{4.45}$$

Similarly to the above, we can derive that

$$\mathbf{G}_{Z_t|Z_{t-1}} = \mathbf{B}_{Z_t|X_t^*} \mathbf{C}_{X_t^*|Z_{t-1}}$$

where $\mathbf{G}$ is likewise a $K \times K$ matrix, defined as

$$\mathbf{G}_{Z_t|Z_{t-1}} = \left[P_{Z_t|Z_{t-1}}(i|j) \right]_{i,j}. \tag{4.46}$$

From Assumption 4.3.3(i), we combine the two previous matrix equalities to obtain

$$\mathbf{A}_{y,Z_t|Z_{t-1}} \mathbf{G}_{Z_t|Z_{t-1}}^{-1} = \mathbf{B}_{Z_t|X_t^*} \mathbf{D}_{y|X_t^*} \mathbf{B}_{Z_t|X_t^*}^{-1}. \tag{4.47}$$

Since $\mathbf{D}_{y|X_t^*}$ is a diagonal matrix, this equation represents an eigenvalue decomposition of the matrix $\mathbf{A}_{y,Z_t|Z_{t-1}} \mathbf{G}_{Z_t|Z_{t-1}}^{-1}$, which can be computed from the observed data sequence

$\{Y_t, Z_t\}$.[47] This shows that from the observed data, we can identify the matrices $\mathbf{B}_{Z_t|X_t^*}$ and $\mathbf{D}_{y|X_t^*}$, which are the matrices with entries equal to (respectively) the eye movement probabilities $P(Z_t|X_t^*)$ and choice probabilities $P(Y_t|X_t^*)$.

In order for this argument to be valid, the eigendecomposition in Eq. (4.47) must be unique. This requires the eigenvalues (corresponding to choice probabilities $P(y|X_t^*)$) to be distinctive; that is, $P(y|X_t^*)$ should vary in X_t^* – which Assumption 4.3.1 ensures. Furthermore, the eigendecomposition in Eq. (4.47) is invariant to the ordering (or permutation) and scalar normalization of eigenvectors. Assumption 4.3.3(ii) imposes the correct ordering on the eigenvectors: specifically, it implies that columns with higher average value correspond to larger value of X_t^*. Finally, because the eigenvectors correspond to the conditional probabilities $P(Z_t|X_t^*)$, it is appropriate to normalize each column so that it sums to one.

Second step. We begin by factorizing the conditional probability distribution

$$\begin{aligned} P(Z_{t+1}, Y_t, R_t, Z_t) &= \sum_{X_t^*} \sum_{X_{t+1}^*} P(Z_{t+1}|X_{t+1}^*) P(X_{t+1}^*|Y_t, X_t^*, R_t) P(Z_t|X_t^*) f(Y_t, X_t^*, R_t) \\ &= \sum_{X_t^*} \sum_{X_{t+1}^*} P(Z_{t+1}|X_{t+1}^*) P(X_{t+1}^*, Y_t, X_t^*, R_t) P(Z_t|X_t^*) \end{aligned}$$

where the second equality applies assumptions 1, 2, and 3. Then, for any fixed $Y_t = y$ and $R_t = r$, we have the matrix equality

$$\mathbf{H}_{Z_{t+1},y,r,Z_t} = \mathbf{B}_{Z_{t+1}|X_{t+1}^*} \mathbf{L}_{X_{t+1}^*,X_t^*,y,r} \mathbf{B'}_{Z_t|X_t^*}.$$

The $K \times K$ matrices $\mathbf{H}$ and $\mathbf{L}$ are defined as

$$\begin{aligned} \mathbf{H}_{Z_{t+1},y,r,Z_t} &= \left[P_{Z_{t+1},Y_t,R_t,Z_t}(i,y,r,j)\right]_{i,j} \\ \mathbf{L}_{X_{t+1}^*,X_t^*,y,r} &= \left[P_{X_{t+1}^*,X_t^*,Y_t,R_t}(i,j,y,r)\right]_{i,j}. \end{aligned} \tag{4.48}$$

By stationarity (assumption 4.3.4), we have $\mathbf{B}_{Z_{t+1}|X_{t+1}^*} = \mathbf{B}_{Z_t|X_t^*}$. Hence, we can obtain $\mathbf{L}_{X_{t+1}^*,X_t^*,y,r}$ (corresponding to the learning rule probabilities) directly from

$$\mathbf{L}_{X_{t+1}^*,X_t^*,y,r} = \mathbf{B}_{Z_{t+1}|X_{t+1}^*}^{-1} \mathbf{H}_{Z_{t+1},y,r,Z_t} [\mathbf{B'}_{Z_t|X_t^*}]^{-1}. \tag{4.49}$$

This result implies that two consecutive periods of experimental and eye movement data $(Z_t, Y_t, R_t), (Z_{t-1}, Y_{t-1}, R_{t-1})$ from each subject suffice to identify and estimate the decision and learning rules in this model.

Our estimation procedure mimics the two-step identification argument from the previous section. That is, for fixed values of (y, r), we first form the matrices $\mathbf{A}$, $\mathbf{G}$, and $\mathbf{H}$ (as defined previously) from the observed data, using sample frequencies to estimate the corresponding probabilities. Then we obtain the matrices $\mathbf{B}$, $\mathbf{D}$, and $\mathbf{L}$ using the matrix manipulations in Eqs. (4.47) and (4.49). To implement this, we assume that the eye movement measures Z_t and the unobserved beliefs X_t^* are discrete, and take three values.[48]

[47] From Eq. (4.46), the invertibility of $\mathbf{G}$ (which is Assumption 4.3.3(i)) implies the invertibility of $\mathbf{B}$.

[48] The details concerning the discretization of the eye movement measure Z_t are given in Appendix C.

Moreover, while the identification argument above was "cross-sectional" in nature, being based upon two observations of $\{Y_t, Z_t, R_t\}$ per subject, in the estimation we exploited the long time series data we have for each subject, and pooled every two time-contiguous observations $\{Y_{i,r,\tau}, Z_{i,r,\tau}, R_{i,r,\tau}\}_{\tau=t-1}^{\tau=t}$ across all subjects i, all blocks r, and all trials $\tau = 2, \ldots, 25$.[49] Results from Monte Carlo simulations (available from the authors on request) show that the estimation procedure produces accurate estimates of the model components.[50]

4.3.3 Results

Tables 4.4 and 4.5 present estimation results. The beliefs X_t^* are assumed to take the three values $\{1, 2, 3\}$. We interpret $X^* = 1, 3$ as indicative of "strong beliefs" favoring (respectively) green and blue, while the intermediate value $X^* = 2$ indicates that the subject is "not sure".[51] Accordingly, the eye movements Z_t have been discretized to also take three values, as discussed before.

Table 4.4 contains the estimates of the choice and eye movement probabilities. The first and last columns of the panels in this table indicate that choices and eyes movements are closely aligned with beliefs, when beliefs are sufficiently strong (ie. are equal to either $X^* = 1$ or $X^* = 3$). Specifically, in these results, the probability of choosing a color contrary to beliefs – which is called the "exploration probability" in the literature – is small, being equal to 1.3% when $X_t^* = 1$, and only 0.64% when $X_t^* = 3$.[52]

When $X_t^* = 2$, however, suggesting that the subject is unsure of the state, there is a slight bias in choices towards "blue", with $Y_t = 2$ roughly 56% of the time. The bottom panel indicates that when subjects are not sure, they tend to split their gaze more evenly between the two colors (ie. $Z_t = 2$) around 63% of the time.

The learning rule estimates are presented in Table 4.5. The left columns show how beliefs are updated when "exploitative" choices (ie. choices made in accordance with beliefs) are taken, and illustrate an important asymmetry in subjects' belief-updating rules. When current beliefs indicate "green" ($X_1^* = 1$) and green is chosen ($Y_t = 1$), beliefs evolve asymmetrically depending on the reward: if $R_t = 2$ (high reward), then beliefs update

[49] Formally, this is justified under the assumption that the process $\{Y_t, Z_t, R_t\}$ is stationary and ergodic for each subject and each block; under these assumptions, the ergodic theorem ensures that the (across time and subjects) sample frequencies used to construct the matrices **A**, **G**, and **H** converge towards population counterparts.

[50] Moreover, because all the elements in the matrices of interest **B**, **D**, and **L** correspond to probabilities, they must take values within the unit interval. However, in the actual estimation, we found that occasionally the estimates do go outside this range. In these cases, we obtained the estimates by a least-squares fitting procedure, where we minimized the elementwise sum-of-squares corresponding to Eqs. (4.47) and (4.49), and explicitly restricted each element of the matrices to lie $\in [0, 1]$. This was not a frequent recourse; only a handful of the estimates reported below needed to be restricted in this manner.

[51] We have tried to re-estimate the model allowing for more belief states (≥ 4), but the results we obtained were not encouraging. This is due to our relatively small sample size; since our estimation approach is nonparametric, it is difficult to obtain reliable estimates with modest sample sizes.

[52] We also considered a robustness check against the possibility that subjects' gazes immediately before making their choices coincide exactly with their choice. While this is not likely in our experimental setting, because subjects were required to indicate their choice by pressing a key on the keyboard, rather than clicking on the screen using a mouse, we nevertheless re-estimated the models but eliminating the last segment of the reaction time in computing the Z_t. The results are very similar to the reported results, both qualitatively and quantitatively.

Table 4.4: Estimates of choice and eye movement probabilities

Estimated Choice Probabilities: $P(Y_t|X_t^*)$

X_t^*:	1(green)	2(not sure)	3(blue)
$Y_t = 1$ (green)	0.9866 (0.0561)	0.4421 (0.1274)	0.0064 (0.0146)
2 (blue)	0.0134	0.5579	0.9936

Estimated eye movement probabilities: $P(Z_t|X_t^*)$

X_t^*:	1(green)	2(not sure)	3(blue)
$Z_t = 1$ (green)	0.8639 (0.0468)	0.2189 (0.1039)	0.0599 (0.0218)
2 (middle)	0.0815 (0.0972)	0.6311 (0.1410)	0.0980 (0.0369)
3 (blue)	0.0546 (0.0581)	0.1499 (0.1206)	0.8421 (0.0529)

Each cell contains parameter estimates, with bootstrapped standard errors in parentheses. Each column sums to one.

towards green with probability 89%; however, if $R_t = 1$ (low reward), then belief still stay at green with probability 57%. This tendency of subjects to update up after successes, but not update down after failures also holds after a choice of "blue" (as shown in the left-hand columns of the bottom two panels in Table 4.5): there, subjects update their belief on blue up to 88% following a success ($R_t = 2$), but still give the event blue a probability of 53% following a failure ($R_1 = 1$). This muted updating following failures is a distinctive feature of our learning rule estimates and, as we will see below, is at odds with optimal Bayesian belief-updating.

The results in the right-most columns describe belief updating following "explorative" (contrarian to current beliefs) choices. For instance, considering the top two panels, when current beliefs are favorable to "blue" ($X_t^* = 3$), but "green" is chosen, beliefs update more towards "green" ($X_{t+1}^* = 1$) after a low rather than high reward (82% vs. 18%). However, the standard errors (computed by bootstrap) of the estimates here are much higher than the estimates in the left-hand columns; this is not surprising, as the choice probability estimates in Figure 4.4 show that explorative choices occur with very low probability, leading to imprecision in the estimates of belief-updating rules following such choices.

The middle columns in these panels show how beliefs evolve following (almost-) random choices. Again considering the top two panels, we see that when current beliefs are unsure ($X_t^* = 2$), subjects update more towards "green" when a previous choice of green yielded the high rather than the low reward (66% vs. 31%). The results in the bottom two panels are very similar to those in the top two panels, but describe how subjects update beliefs following choices of "blue" ($Y_t = 2$).

Table 4.5: Estimates of learning (belief-updating) rules

$P(X^*_{t+1}|X^*_t, y, r)$, r =1(lose), y =1(green)

X^*_t:	1(green)	2 (not sure)	3(blue)
$X^*_{t+1}=1$ (green)	0.5724 (0.0694)	0.3075 (0.0881)	0.1779 (0.2257)
2 (not sure)	0.0000[a] (0.0662)	0.3138 (0.1042)	0.4002 (0.2284)
3 (blue)	0.4276 (0.0624)	0.3787 (0.0945)	0.4219 (0.2195)

$P(X^*_{t+1}|X^*_t, y, r)$, r =2(win), y =1(green)

X^*_t:	1(green)	2 (not sure)	3(blue)
$X^*_{t+1}=1$ (green)	0.8889 (0.0894)	0.6621 (0.1309)	0.8242 (0.2734)
2 (not sure)	0.0000 (0.0911)	0.2702 (0.1297)	0.1758 (0.1981)
3 (blue)	0.1111 (0.0340)	0.0678 (0.0485)	0.0000 (0.1876)

$P(X^*_{t+1}|X^*_t, y, r)$, r =1(lose), y =2(blue)

X^*_t:	3(blue)	2 (not sure)	1(green)
$X^*_{t+1}=3$ (blue)	0.5376 (0.0890)	0.2297 (0.0731)	0.2123 (0.1436)
2 (not sure)	0.0458 (0.0732)	0.2096 (0.0958)	0.1086 (0.1524)
1 (green)	0.4166 (0.0874)	0.5607 (0.0968)	0.6792 (0.1881)

$P(X^*_{t+1}|X^*_t, y, r)$, r =2(win), y =2(blue)

X^*_t:	3(blue)	2 (not sure)	1(green)
$X^*_{t+1}=3$ (blue)	0.8845 (0.1000)	0.6163 (0.1136)	0.6319 (0.1647)
2 (not sure)	0.0000 (0.0968)	0.3558 (0.1160)	0.3566 (0.1637)
1 (green)	0.1155 (0.0499)	0.0279 (0.0373)	0.0116 (0.0679)

Each cell contains parameter estimates, with bootstrapped standard errors in parentheses.
Each column sums to one.

[a]This estimate, as well as the other estimates in this table which are equal to zero, resulted from applying the constraint that probabilities must lie between 0 and 1. See footnote 50.

Table 4.6: Simulated payoffs from learning models

	Fully-rational Bayesian	Nonparametric	Pseudo-Bayesian	Reinforcement Learning	Win-stay
5-%tile	\$5	\$1	\$2	\$1	\$1
25-%tile	\$12	\$8	\$9	\$8	\$8
50-%tile	\$17	\$13	\$14	\$13	\$13
75-%tile	\$22	\$18	\$19	\$18	\$18
95-%tile	\$29	\$25	\$26	\$25	\$25

The fully-rational model is described in Section 1.1, while the Reinforcement learning, Pseudo-Bayesian, and win-stay models are described in Appendix B. For each model, the quantiles of the simulated payoff distribution (across 100,000 simulated choice/reward sequences) are reported.

4.3.4 How Optimal are Estimated Learning Rules

In the remainder of the paper, we compare our estimated learning rules to alternative learning rules which have been considered in the literature. We consider four alternative parametric learning rules: (i) the *fully-rational Bayesian* model, which is the model discussed in Section 1.1 above; (ii) a *pseudo-Bayesian* model, which is a version of the fully-rational Bayesian model in which the decision rules are smoothed relative to the step-function decision rules in the fully-rational model ; (iii) *reinforcement learning* (cf. Sutton and Barto (1998)); and (iv) *win-stay*, a simple choice heuristic whereby subjects replay successful strategies. All of these models, except (i), contain unknown model parameters, which we estimated using the choice data from the experiments. Complete details on these models, and the estimated model parameters, are given in Appendix B.

The relative optimality of each learning model was assessed via simulation. For each model, we simulated 100,000 sequences (each containing eight blocks of choices, as in the experiments) of rewards and choices, and computed the distributions of payoffs obtained by subjects. The empirical quantiles of these distributions are presented in Table 4.6.

As we expect, the fully-rational Bayesian model generates the most revenue for subjects; the payoff distribution for this model stochastically dominates the other models, and the median payoff is \$17. The other models perform almost identically, with a median payoff around \$3-\$4 less than the Bayesian model (or about two cents per choice). This difference accounts for about 25% of typical experimental earnings (net of the fixed show-up fee).

In the next section, we seek explanations for the differences (and similarities) in performance among the alternative learning models by comparing the belief-updating and choice rules across the different models.

Comparing Choice and Belief-Updating Rules Across Different Learning Models

For the fully-rational Bayesian and reinforcement learning models, we can recover the "beliefs" corresponding to the observed choices and rewards, and compare them to the beliefs from the nonparametric learning model.[53] Appendix B contains additional details on how

[53]There are no beliefs in the win-stay model, which is a simple choice heuristic. The pseudo-Bayesian model has the same beliefs as the fully-rational Bayesian model (with the difference that the choice rule is

Table 4.7: Summary statistics for beliefs in three learning models

X^*: Beliefs from nonparametric model
B^*: Beliefs from fully-rational Bayesian model
V^*: "Beliefs" (valuations) from reinforcement learning model

Panel 1: Belief frequency in nonparametric model

X^*	1(green)	2(not sure)	3(blue)
	1878 (45%)	366 (10%)	1956 (45%)

Panel 2: Beliefs from other models

	mean	median	std.	lower 33%	upper 33%
B^* (Bayesian Belief)	0.4960	0.5000	0.1433	0.4201	0.5644
$V^*(= V_b - V_g)$	-0.0104	0	0.4037	-0.2095	0.1694

All three measures of beliefs are oriented so that larger values correspond to a more favorable assessment that "blue" is currently the good arm.

See Appendix B for details on computation of beliefs in these three learning models.

the beliefs were derived for the learning models.

Table 4.7 contains summary statistics for the implied beliefs from our nonparametric learning model (denoted X_t^*), vs. the Bayesian beliefs B^* and the valuations V^* in the Reinforcement Learning model. For simplicity, we will abuse terminology somewhat and refer in what follows to X^*, V^*, and B^* as the "beliefs" implied by, respectively, our nonparametric model, the Reinforcement Learning model, and the Bayesian model.

Panel 1 gives the total tally, across all subjects, blocks, and trials, of the number of times the nonparametric beliefs X^* took each of the three values. Subjects' beliefs tended to favor green and blue roughly equally, with "not sure" lagging far behind. The close split between "green" and "blue" beliefs is consistent with the notion that subjects have rational expectations, with flat priors on the unobserved state S_1 at the beginning of each block. The second panel summarizes the beliefs from the Reinforcement Learning and Bayesian models. The Reinforcement Learning valuation measure V^* appears symmetric and centered around zero, while the average Bayesian belief B^* lies also around 0.5. Thus, on the whole, all three measures of beliefs appear equally distributed between "green" and "blue".

Next, we compare the learning rules from the nonparametric, fully-rational Bayesian, and reinforcement learning models. In order to do this, we discretized the beliefs in each model into three values, in proportions identical to the frequency of the different values of X_t^* as reported in Table 4.7, and present the implied learning rules for each model.[54] These are shown in Table 4.8.

smoothed).

[54]Specifically, we discretized the Bayesian (resp. Reinforcement Learning) beliefs so that 45% of the beliefs fell in the $B_t^* = 1$(resp. $V_t^* = 1$) and $B_{t+1}^* = 3$(resp. $V_t^* = 3$) categories, while 10% fell in the intermediate $B_t^* = 2(X_t^* = 2)$ category, the same as for the nonparametric beliefs X_t^* (cf. Panel 1 of Table 6). The results are even more striking when we discretized the Bayesian and Reinforcement Learning beliefs so that 33% fell into each of the three categories.

Table 4.8: Learning (belief-updating) rules for alternative learning models

$P(X^*_{t+1}|X^*_t, y, r)$, r =1(lose), y =1(green)

	Fully-rational Bayesian			**Reinforcement Learning**		
Beliefs B^*_{t+1}, V^*_{t+1}:	1(green)	2 (not sure)	3(blue)	1(green)	2 (not sure)	3(blue)
1 (green)	0.2878	0	0	0.6538	0	0
2 (not sure)	0.1730	0	0	0.1381	0.0115	0
3 (blue)	0.5392	1.0000	1.0000	0.2080	0.9885	1.0000

$P(X^*_{t+1}|X^*_t, y, r)$, r =2(win), y =1(green)

	Fully-rational Bayesian			**Reinforcement Learning**		
Beliefs B^*_{t+1}, V^*_{t+1}:	1(green)	2 (not sure)	3(blue)	1(green)	2 (not sure)	3(blue)
1 (green)	1.0000	1.0000	0.6734	1.0000	0.8818	0.6652
2 (not sure)	0	0	0.1250	0	0.1182	0.1674
3 (blue)	0	0	0.2016	0	0	0.1674

$P(X^*_{t+1}|X^*_t, y, r)$, r =1(lose), y =2(blue)

	Fully-rational Bayesian			**Reinforcement Learning**		
Beliefs B^*_{t+1}, V^*_{t+1}:	3(blue)	2 (not sure)	1(green)	3(blue)	2 (not sure)	1(green)
3 (blue)	0.3060	0	0	0.6576	0	0
2 (not sure)	0.1601	0	0	0.1261	0.0109	0
1 (green)	0.5338	1.0000	1.0000	0.2164	0.9891	1.0000

$P(X^*_{t+1}|X^*_t, y, r)$, r =2(win), y =2(blue)

	Fully-rational Bayesian			**Reinforcement Learning**		
Beliefs B^*_{t+1}, V^*_{t+1}:	3(blue)	2 (not sure)	1(green)	3(blue)	2 (not sure)	1(green)
3 (blue)	1.0000	1.0000	0.6760	1.0000	0.8898	0.6983
2 (not sure)	0	0.0000	0.1440	0	0.1102	0.1379
1 (green)	0	0	0.1800	0	0	0.1638

All three measures of beliefs are oriented so that larger values correspond to a more favorable assessment that "blue" is currently the good arm.

Table 4.9: Choice probabilities for alternative learning models

Fully-rational Bayesian			
Beliefs B_t^*:	1(green)	2(not sure)	3(blue)
$Y_t = 1$ (green)	1.0000	0.5000	0.0000
2 (blue)	0.0000	0.5000	1.0000

Pseudo-Bayesian Learning			
Beliefs B_t^*:	1(green)	2(not sure)	3(blue)
$Y_t = 1$ (green)	0.5141	0.4996	0.4850
2 (blue)	0.4859	0.5005	0.5150

Reinforcement Learning			
Beliefs V_t^*:	1(green)	2(not sure)	3(blue)
$Y_t = 1$ (green)	0.7629	0.4939	0.2250
2 (blue)	0.2371	0.5061	0.7750

All three measures of beliefs are oriented so that larger values correspond to a more favorable assessment that "blue" is currently the good arm.

The most striking difference between the three sets of learning rules lies in how beliefs update following unsuccessful choices (ie. choices which yielded a negative reward). Comparing the Bayesian and the nonparametric learning rules (in Table 4), we see that Bayesian beliefs exhibit less "stickiness", or serial correlation, following unsuccessful choices. For example, consider the case of $(Y_t = 1, R_t = 1)$, so that an unsuccessful choice of green occurred in the previous period. The nonparametric learning rule estimates (Table 4) show that the weight of beliefs remain on green $(X_{t+1}^* = 1)$ with 57% probability, whereas the Bayesian beliefs place only 28% weight on green. A similar pattern exists after an unsuccessful choice of blue, as shown in the left-hand column of the third panel: the nonparametric learning rule continues to place 54% probability on blue, whereas the fully-rational Bayesian belief is only 30%.

On the other hand, the learning rules for the Reinforcement Learning model (also reported in Table 4.8) are more similar to the nonparametric learning rule, especially following unsuccessful choices. Again, looking at the top panel, we see that following an unsuccessful choice of "green" $(Y_t = 1)$, subjects' valuations are still favorable to green with probability 65%; this is comparable in magnitude to the 57% from the nonparametric learning rule. Similarly, after an unsuccessful choice of blue (third panel), valuations in the Reinforcement Learning model still favor blue with probability 66%, again comparable to the 54% for the nonparametric model. It appears, then, that the updating rules from the Reinforcement Learning and nonparametric model share a common defect: a reluctance to "update down" following unsuccessful choices; this common defect relative to the fully-rational model may explain the lower revenue generated by these models.

In Table 4.9 we compare the choice rules across the different models. As in the previous table, we discretized the beliefs from each model into three values. Comparing the top two panels, we see that, even though the belief-updating rule is the same for the fully-rational

Bayesian and Pseudo-Bayesian models, the choice rules are strikingly different. Choice rules in the fully-rational Bayesian model are binary deterministic functions of beliefs . In contrast, the Pseudo-Bayesian model is a model in which the choice rule is a "smoothed" probabilistic function of beliefs; the estimate of the smoothing parameter (discussed in Appendix B) indicates a large amount of smoothing, such that choice probabilities in this model are practically invariant to the Bayesian beliefs B_t^*, as shown in Table 4.9.

Overall, the estimated choice rules for the nonparametric model, in Table 4.4, are much closer to the fully-rational model, than the Pseudo-Bayesian model. This suggests that the lower payoffs from the nonparametric model relative to the fully-rational model arise primarily not from the choice rules (which are very similar in the two models), but rather from the belief-updating rules (which are quite different, as discussed previously).

The bottom panel of Table 4.9 contains the choice rules for the Reinforcement Learning model. As shown there, the choice rules are much smoother than in the fully-rational Bayesian model and the estimated model, but not as smooth as the Pseudo-Bayesian model. This suggests that the similarities of the payoffs from the nonparametric and Reinforcement Learning models (as shown in Table 4.6) are due to the similarities in belief-updating rules, and not to the choice rules, which are quite different in the two models.

Finally, the similarity in payoffs between the nonparametric and win-stay models is not surprising because, as we showed in Section 1.3 above, the reduced-form choice behavior from the experimental data is in line with a "win-stay/lose-randomize" rule of thumb. Such behavior is confirmed in the calibrated parameters for the win-stay model (presented in Appendix B) which show that, after receiving a positive reward, subjects tend to repeat the previous choice with probability 87% while, after a negative reward, subjects essentially randomize. This asymmetry in choices following good/bad rewards echoes the estimated learning rules from Table 4, which showed that subjects "update down" much less following bad rewards than they "update up" following good rewards.

4.3.5 Additional Details

Details for Computing Fully-Rational Bayesian Learning Model

Here we provide more details about the simulation of the fully-rational model from Section 2.1. First we introduce some notation and describe the information structure and how Bayesian updating would proceed in the reversal learning context. Let (Y_t, S_t, R_t) denote the actions, state, and rewards. Furthermore, let Q denote the 2×2 Markov transition matrix for the state S_t, corresponding to Eq. (2).

Let B_t^* denote the *prior belief* that $S_t = 2$, at the beginning of period t, while $\tilde{B}_t^*$ denotes the *posterior* belief that $S_t = 2$, at the end of period t, after taking action Y_t and observing reward R_t. The relationship between B_t^* and $\tilde{B}_t^*$ is given by Baye's rule:

$$\tilde{p}_t = P(S_t = 2|p_t, R_t, Y_t) = \frac{p_t \cdot P(R_t|S_t = 2, Y_t)}{(1 - p_t) \cdot P(R_t|S_t = 1, Y_t) + p_t \cdot P(R_t|S_t = 2, Y_t)}$$

Combining this with Q, we obtain the period-by-period transition for the prior beliefs B_t^*:

$$\begin{bmatrix} 1 - B_{t+1}^* \\ B_{t+1}^* \end{bmatrix} = Q \cdot \begin{bmatrix} 1 - \tilde{B}_t^* \\ \tilde{B}_t^* \end{bmatrix} = Q \cdot \begin{bmatrix} 1 - P(S_t = 2|B_t^*, R_t, Y_t) \\ P(S_t = 2|B_t^*, R_t, Y_t) \end{bmatrix} \tag{4.50}$$

Next we describe a dynamic Bayesian learning model for the reversal-learning environment. As in the experiments, we consider a finite (25 period) horizon, with $t = 1, \ldots T = 25$. Each subject's objective is to choose sequence of actions to maximize expected rewards:

$$\max_{i_1, i_2, \ldots, i_T} \mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$$

The state variable in this model is B_t^*, the beliefs at the beginning of each period. Correspondingly, the Bellman equation is:

$$\begin{aligned} V_t(B_t^*) &= \max_{Y_t \in \{1,2\}} \left\{\mathbb{E}\left[R_t + V_{t+1}(B_{t+1}^*)|Y_t, B_t^*\right]\right\} \\ &= \max_{Y_t \in \{1,2\}} \left\{\mathbb{E}\left[R_t|Y_t, B_t^*\right] + \mathbb{E}_{R_t|Y_t, B_t^*} \mathbb{E}_{B_{t+1}^*|B_t^*, Y_t, R_t} V_{t+1}(B_{t+1}^*)\right\} \end{aligned} \tag{4.51}$$

Above, the expectation $E_{B_{t+1}^*|B_t^*, Y_t, R_t}$ is taken with respect to Eq. (4.50), the law of motion for the prior beliefs, while the expectation $E_{R_t|Y_t, B_t^*}$ is derived from the assumed distribution of $(R_t|Y_t, \omega_t)$ via

$$P(R_t|Y_t, B_t^*) = (1 - B_t^*) \cdot P(R_t|Y_t, \omega_t = 1) + B_t^* \cdot P(R_t|Y_t, \omega_t = 2).$$

While we have not been able to derive closed-form solutions to this dynamic optimization problem, we can compute the optimal decision rules by backward induction. Specifically, in the last period $T = 25$, the Bellman equation is:

$$V_T(B_T^*) = \max_{Y_t \in \{1,2\}} E\left[R_t|Y_t, B_T^*\right]. \tag{4.52}$$

We can discretize the values of B_T^* into the finite discrete set $\mathcal{B}$. Then for each $B \in \mathcal{B}$, we can solve Eq. (4.52) to obtain the period-T value and choice functions $\hat{V}_T(B)$ and $\hat{y}_T^*(B) = \text{argmax}_i \mathbb{E}[R_t|i, B]$ for each value of $B \in \mathcal{B}$. Subsequently, proceeding backwards, we can obtain the value and choice functions for periods $t = T - 1, T - 2, \ldots, 1$.

Details on Model Fitting and Belief Estimation in Alternative Learning Models

In section 4.3.4, we compared belief dynamics in the nonparametric model (X^*) with counterparts in other dynamic choice models. Here we provide additional details on how these quantities were computing for each model.

Recovering belief dynamics X^* in the nonparametric model. The values of X^*, the belief process in our nonparametric learning model, were obtained by maximum likelihood. For each block, using the estimated choice and eye movement probabilities, as well as the

learning rules, we chose the path of beliefs $\{X_t^*\}_{t=1}^{25}$ which maximized $P(\{X_t^*\} \mid \{Z_t, R_t\})$, the conditional ("posterior") probability of the beliefs, given the observed sequences of eye-movements and rewards. Because

$$P(\{X_t^*, Z_t\} \mid \{Y_t, R_t\}) = P(\{X_t^*\} \mid \{Z_t, R_t\}) \cdot P(\{Z_t\} \mid \{Y_t, R_t\}),$$

where the second term on the RHS of the equation above does not depend on X_t^*, it is equivalent to maximize $P(\{X_t^*, Z_t\} \mid \{Y_t, R_t\})$ with respect to $\{X_t^*\}$. Because of the Markov structure, the joint log-likelihood factors as:

$$\log L(\{X_t^*, Z_t\}|\{Y_t, R_t\}) = \sum_{t=1}^{24} \log \left[P(Z_t|X_t^*)P(X_{t+1}^*|X_t^*, R_t, Y_t)\right] + \log(P(Z_{25}|X_{25}^*)). \tag{4.53}$$

We plug in our nonparametric estimates of $P(Z|X^*)$ and $P(X_{t+1}^*|X_t^*, R_t, Y_t)$ into the above likelihood, and optimize it over all paths of $\{X_t^*\}_{t=1}^{25}$ with the initial condition restriction $X_1^* = 2$ (beliefs indicate "not sure" at the beginning of each block). To facilitate this optimization problem, we derive the best-fitting sequence of beliefs using a dynamic-programming (Viterbi) algorithm; cf. Ghahramani (2001).

In the above, we treated the choice sequence $\{Y_t\}$ as exogenous, and left the choice probabilities $P(Y_t|X_t^*)$ out of the log-likelihood function (4.53) above. By doing this, we essentially ignore the implied correlation between beliefs and choices in estimating beliefs. This was because, given our estimates that $P(Y_t = 1|X_t^* = 1) \approx P(Y_t = 2|X_t^* = 3) \approx 1$ in Table 4.4, maximizing with respect to these choice probabilities would leads to estimates of beliefs $\{X^*\}$ which closely coincide with observed choices; we wished to avoid such an artificially good "fit" between the beliefs and observed choices.

For robustness, however, we also estimated the beliefs $\{X^*\}$ including the choice probabilities $P(Y_t|X_t^*)$ in the likelihood function. Not surprisingly, the correlation between choices and beliefs $\text{Corr}(Y_t, X_t^*) = 0.99$, and in practically all periods, the estimated beliefs and observed choices coincided (ie. $X_t^* = Y_t$). However, we felt that this did not accurately reflect subjects' beliefs.

Beliefs in fully-rational Bayesian Model. The learning and decision rules for the Bayesian model were described and computed in Section 1.1, with additional details provided in Appendix A. The sequence of Bayesian beliefs B_t^* is obtained from Eq. (4.50) and evaluated at the observed sequence of choices and rewards (Y_t, R_t).

Reinforcement Learning Model. We employ a variant of the TD (Temporal-Difference)-Learning models (Sutton and Barto (1998) , section 6) in which action values are up-dated via the so-called Recorla-Wagner rule. The value updating rule for a one-step TD-Learning model is given by:

$$V_{Y_t}^{t+1} \longleftarrow V_{Y_t}^t + \alpha\delta_t. \tag{4.54}$$

where Y_t denotes the choice taken in trial t, α denotes the learning rate, and δ_t denotes the "prediction error" δ_t for trial t, defined as:

$$\delta_t = R_t - V^t_{Y_t}, \tag{4.55}$$

the difference between R_t (the observed reward in trial t) and $V^t_{Y_t}$ (the current valuation). In trial t, only the value for the chosen alternative Y_t is updated; there is no updating of the valuation for the choice that was not taken.

P^t_c, the current probability of choosing action c, is assumed to take the conventional "softmax" (ie. logit) form with the smoothing (or "temperature") parameter τ:

$$P^t_c = e^{V^t_c/\tau} / \left[\sum_{c'} e^{V^t_{c'}/\tau}\right] \tag{4.56}$$

We estimated the parameters τ and α using maximum likelihood. For greater model flexibility, we allowed the parameter α to differ following positive vs. negative rewards. The estimates (and standard errors) are:

$$\begin{aligned} &\tau = 0.2729 \quad (0.0307) \\ &\alpha \text{ for positive reward } (R_t = 2) = 0.7549 \quad (0.0758) \\ &\alpha \text{ for negative reward } (R_t = 1) = 0.3333 \quad (0.0518). \end{aligned} \tag{4.57}$$

We plug in these values into Eqs. (4.54), (4.55), and (4.59) to derive a sequence of valuations $\left\{V^*_t \equiv V^t_b - V^t_g\right\}$. The choice function (Eq. (4.56)) can be rewritten as a function of the difference V^*_t; i.e. the choice probability for the blue slot machine is,

$$P^t_b = \frac{e^{(V^t_b - V^t_g)/\tau}}{1 + e^{(V^t_b - V^t_g)/\tau}} = \frac{e^{V^*_t}/\tau}{1 + e^{V^*_t/\tau}} \tag{4.58}$$

and $P^t_g = 1 - P^t_b$. Hence, V^*_t plays a role in the TD-Learning model analogous to the belief measures X^*_t and B^*_t from, respectively, the nonparametric and Bayesian learning models.

Pseudo-Bayesian Model. A Pseudo-Bayesian learner uses Bayes rule to update her belief (as in the fully-rational model), but her choice probabilities are determined (suboptimally) by the "softmax" rule, as in reinforcement learning:

$$P^t_c = e^{B^{*t}_c/\tau} / \left[\sum_{c'} e^{B^{*t}_{c'}/\tau}\right] \tag{4.59}$$

As the smoothing parameter $\tau \to 0$, the Pseudo-Bayesian model approaches the fully-rational Bayesian model. Using the choice data form the experiments, we obtain a maximum-likelihood estimate of 0.2176 for τ, with a bootstrapped standard error of 0.0138. This indicates a large degree of smoothing in the choice probabilities relative to the fully-rational decision rule, as shown in Table 4.9.

Win-Stay Model. The final model is a simple behavioral heuristic. If subjects choose a slot machine and receive the positive reward $R_t = 1$, they repeat the choice in the next period with probability $1 - \delta$ (and switch to the other choice with probability δ). If they choose a slot machine but obtain the negative reward $R_t = -1$, they switch to the other slot machine in the next trial with probability $1 - \epsilon$.

We estimated the parameters δ and ϵ using maximum likelihood. The estimates we obtained from the data were:

$$\delta = 0.1268 \quad (0.0142); \quad \epsilon = 0.4994 \quad (0.0213). \tag{4.60}$$

4.4 Effort and Types in Online Credit Market

Xin (2018) studies the impact of reputation/feedback systems on the operation of online credit markets using data from Prosper.com. A major concern in markets of unsecured loans is the ability of lenders to recover their loan due to the problems of asymmetric information. On the one hand, borrowers differ in their inherent default costs c, which is hidden information; on the other hand, borrowers' efforts e_t to repay debts are hidden as well, so additional incentives are necessary to motivate them.

Xin (2018) is the first to quantify the extent to which reputation/feedback systems improve the total welfare of market participants when both hidden information (adverse selection) and hidden actions (moral hazard) are present. She identifies and estimates a finite-horizon dynamic model of a credit market in which borrowers and lenders interact repeatedly over time. The observables include the outcome variables O_t, including default and late payment performances, and individual characteristics X_t, such as debt-to-income ratio and credit grade.

The dynamic structure implies that

$$f(O_t, X_t, O_{t-1}, X_{t-1}) = \sum_c f(c, X_{t-1}, O_{t-1}) f(X_t | X_{t-1}, O_{t-1}, c) f(O_t | c, X_t)$$

The type distribution $f(c|X_{t-1}, O_{t-1})$ is identified for borrowers with multiple loans using the identification results in Hu and Shum (2012).

Furthermore, loan outcomes include borrowers' default and late payment performances, $O_t = \{D_t, L_t\}$. The model implies that default and late payment are independent conditional on effort, i.e.,

$$f(O_t | c, X_t) = \sum_{e_t} f(D_t | e_t) f(L_t | e_t) f(e_t | c, X_t)$$

Following the results in Hu (2008), effort choice probabilities and outcome realization process are identified.

These results lead to identification of utility parameters in borrowers' payoff functions and the outside option distributions for borrowers and lenders using variations in interest rates. In the last step, given other primitives that have been recovered, she identifies

the original type distribution for all borrowers before any selection occurs. Using these structural estimates, the paper also conducts counterfactual experiments.

More detailed description can be found in Yi Xin's presentation slides ↗.

5

Applications in Labor Economics

5.1 Unemployment and Labor Force Participation

Unemployment rates may be one of the most important economic indicators. The official US unemployment rates are estimated using self-reported labor force statuses in the Current Population Survey (CPS). It is known that ignoring misreporting errors in the CPS may lead to biased estimates. Feng and Hu (2013) use a hidden Markov approach to identify and estimate the distribution of the true labor force status. Let X_t^* and X_t denote the true and self-reported labor force status in period t. They merge monthly CPS surveys and are able to obtain a random sample $\{X_{t+1}, X_t, X_{t-9}\}_i$ for $i = 1, 2, \ldots, N$. Using X_{t-9} instead of X_{t-1} may provide more variation in the observed labor force status. They assume that the misreporting error only depends on the true labor force status in the current period, and therefore,

$$\begin{aligned} & \Pr\left(X_{t+1}, X_t, X_{t-9}\right) \\ = & \sum_{X_{t+1}^*} \sum_{X_t^*} \sum_{X_{t-9}^*} \Pr\left(X_{t+1}|X_{t+1}^*\right) \Pr\left(X_t|X_t^*\right) \Pr\left(X_{t-9}|X_{t-9}^*\right) \Pr\left(X_{t+1}^*, X_t^*, X_{t-9}^*\right). \end{aligned} \tag{5.1}$$

With three unobservables and three observables, nonparametric identification is not feasible without further restrictions. They then assume that $\Pr\left(X_{t+1}^*|X_t^*, X_{t-9}^*\right) = \Pr\left(X_{t+1}^*|X_t^*\right)$, which is similar to a first-order Markov condition. Under these assumptions, they obtain

$$\begin{aligned} & \Pr\left(X_{t+1}, X_t, X_{t-9}\right) \\ = & \sum_{X_t^*} \Pr\left(X_{t+1}|X_t^*\right) \Pr\left(X_t|X_t^*\right) \Pr\left(X_t^*, X_{t-9}\right), \end{aligned} \tag{5.2}$$

which implies a 3-measurement model. This model can be considered as an application of Theorem 2.4.1 to a hidden Markov model.

Feng and Hu (2013) found that the official U.S. unemployment rates substantially underestimate the true level of unemployment, due to misreporting errors in the labor force status in the Current Population Survey. From January 1996 to August 2011, the corrected monthly unemployment rates are 2.1 percentage points higher than the official rates on

average, and are more sensitive to changes in business cycles. The labor force participation rates, however, are not affected by this correction.

5.1.1 Background

The unemployment rate is among the most important and carefully-watched economic indicators in modern society, and often takes center stage in discussions of economic policy. However, there is considerable disagreement over the precise definition and measurement of unemployment, hence the other two labor force statuses: "employed" and "not-in-labor-force".[1] In the United States, the Bureau of Labor Statistics (BLS) reports six alternative measures of unemployment (U1-U6), including the official unemployment rate (U3) which is based on the International Labor Organization (ILO)'s definition.[2] Due to the intrinsic difficulties in classifying some groups of people, such as marginally-attached workers and involuntary part-time workers, into three distinct labor force statuses, the U.S. official unemployment rate is potentially subject to measurement error.

In this paper, we take a latent variables approach and view the reported labor force statuses as functions of the underlying unobserved true labor force statuses. We then impose a structure on the misclassification process and the dynamics of the underlying latent LFS. Using recent results in the measurement error literature, we show that the official U.S. unemployment rate substantially underestimates the true level of unemployment. During the period from January 1996 to August 2011, our corrected unemployment rates are higher than the corresponding official figures by 2.1 percentage points on average. In terms of the monthly differences, the corrected rates are from 1 to 4.4 percentage points higher than the official rates, and are more sensitive to changes in the business cycles.

Official unemployment statistics in U.S. are based on the Current Population Survey (CPS) conducted by the Census Bureau. The CPS interviews around 60,000 households each month to collect basic demographic and labor force status information. Based on the answers to survey questions on job-related activities, the CPS records each individual's labor force status as "employed", "unemployed," or "not-in-labor-force." The misclassification among the three possible values of the labor force status has been a substantial issue in the CPS, as clearly demonstrated by the Reinterview Surveys, in which a small sub-sample of the households included in the original CPS are recontacted and asked the same questions. Treating the CPS reconciled Reinterview Surveys sample as reflecting true labor force statuses, researchers have found that there exists considerable error in the original CPS.[3] Of course, the actual misclassification errors in labor force status are likely to be substantially

[1] For example, using Canadian data, Jones and Riddell (1999) empirically examine labor market transitions of people with different labor force statuses and find substantial heterogeneity within the nonemployed, such that no dichotomy exists between those unemployed and not-in-labor-force among all nonemployed persons.

[2] The ILO defines "unemployed" as those who are currently not working but are willing and able to work for pay, currently available to work, and have actively searched for work.

[3] The CPS reinterview sample consists two components, one is "non-reconciled", in which case no attempt is made to determine which answers are correct, the other is "reconciled", in which case the second interviewer would compare the responses from the first survey with the reinterview answers and try to resolve any conflicts (Poterba and Summers, 1984).

larger than suggested in reconciled CPS reinterviews, as argued by Poterba and Summers (1995), Biemer and Forsman (1992) and Sinclair and Gastwirth (1996).

The misclassification of labor force statuses in the CPS and other similar surveys has received considerable attention in the literature. To identify the misclassification probabilities, early studies typically relied on a particular exogenous sources of "truth", such as the reconciled CPS reinterview surveys (see e.g. Abowd and Zellner 1985, Poterba and Summers 1986, and Magnac and Visser 1999). Nevertheless, the reinterview surveys are usually small in sample size (approximately 3% of the corresponding CPS sample) and not readily available to outside researchers. Reinterview surveys are also subject to misclassification errors due to many practical limitations.[4] Actually, some studies using other methods show that the reconciled CPS reinterview data may contain even more errors than the original CPS sample (Sinclair and Gastwirth, 1996). Other studies rely on two repeated measures of the labor force status of the same individuals in the same period and assume that the error probabilities are the same for different sub-samples.[5] More recent studies, such as Biemer and Bushery (2000) and Bassi and Trivellato (2008), explore the panel nature of the surveys and treat the underlying true labor force status as a latent process to be jointly modeled with the misclassification process.

Most existing studies focus on adjusting flows, i.e., the gross labor flows between two consecutive months, not stocks, such as the unemployment rate and the labor market participation rate. While those studies acknowledge that misclassification errors cause serious problems for flows, they somewhat surprisingly assume that errors tend to cancel out for stocks (e.g. Singh and Rao 1995). The only study that has tried to correct for the unemployment rate is Sinclair and Gastwirth (1998). However, their results rely on a key identification assumption that males and females have the same misclassification error probabilities, which we reject in this paper.

This paper uses recent results in the measurement error literature to identify the misclassification probabilities (Hu, 2008). Our method relies only on short panels formed by matching the CPS monthly data sets, thus avoiding the use of auxiliary information such as the reinterview surveys, which are usually small and subject to errors. Our approach is close to the Markov Latent Class Analysis (MLCA) method proposed by Biemer and Bushery (2000), but we use an eigenvalue-eigenvector decomposition to establish a closed-form global identification, while they took a maximum likelihood approach with local identifiability. Generally speaking, parametric GMM or MLE methods typically rely on a local identification argument that the number of unknowns does not exceed that of the restrictions. Given the observed distribution, our identification and estimation procedure directly leads to the unique true values of the unknown probabilities without using the

[4]The reinterview may not have been independent of the original interview to the extent that respondents remembered and repeated their answers from the original interview. In addition, several factors make it difficult to conduct the reinterview as an exact replication of the original interview, including (1) Only senior interviewers conducted the reinterview, (2) Almost all reinterviews were conducted by telephone, even if the original interview was conducted in person, and (3) The reinterview may not perfectly "anchor" respondents in the original interview's reference period.

[5]See Sinclair and Gastwirth (1996, 1998), which use the H-W model first proposed by Hui and Walter (1980).

regular optimization algorithms. Therefore, we do not need to be concerned about choosing initial values or obtaining a local maximum in the estimation procedure. In that sense, our estimates are more reliable than those based on local identification, including Biemer and Bushery (2000). Our assumption regarding the dynamics of the underlying true labor force status is also weaker than their first-order Markov chain assumption. In addition, Biemer and Bushery (2000) use group-level data, which are subject to potential biases from within-group heterogeneities. Our identification results enable us to take advantage of the large sample size of the individual-level CPS data, and therefore, to achieve more efficient estimates.

To control for individual heterogeneities, we separately estimate the misclassification probabilities for each demographic group, defined by individual's gender, race and age. Based on those misclassification probabilities, we then estimate the corrected monthly unemployment rates and the labor force participation rates for all demographic groups, and for the US population as a whole. During the period from January 1996 to August 2011, our corrected unemployment rates are higher than the official ones by up to 4.4 percentage points and on average by 2.1 percentage points, with the differences always statistically significant. The most substantial misclassification errors occur when unemployed individuals misreport as either not-in-labor-force or employed. On the other hand, the corrected labor force participation rates and the official ones are rather close and never statistically significantly different.

The rest of the paper is organized as follows. Section 2 provides theoretical results on the identification and estimation of the misclassification probabilities and the marginal distribution of the underlying labor force status. Section 3 presents our main empirical results on the estimated misclassification probabilities and the corrected unemployment rates, along with reported (official) ones. The last section concludes. Additional estimates and simulation results are included in the online appendix of the paper.

5.1.2 A Closed-Form Identification Result

This section presents a closed-form identification and estimation procedure, which uniquely maps the directly estimable distribution of the self-reported labor force status to the misclassification probabilities and the distribution of the underlying true labor force status. We also evaluate the validity and robustness of the assumptions made in order to achieve identification.

Assumptions and Identification Results

Let U_t denote the self-reported labor force status in month t, and X be a vector of demographic variables such as gender, race and age. By matching the monthly CPS samples, we observe the self-reported labor force status in three periods $(t+1, t, t-9)$, together with the demographic variables X for each individual i.[6] For example, if U_t stands for the labor

[6]Our identification strategy requires matching of three CPS monthly data sets in order to identify the misclassification matrix for the month in the middle of the three months. We choose one month later, i.e., $t+1$, and nine month earlier, i.e., $t-9$, for the following reasons: 1) we want the three periods to be

force status of an individual in January 2008, then U_{t+1} and U_{t-9} denote his or her labor force status in February 2008 and in April 2007, respectively. We denote the i.i.d. sample as $\{U_{t+1}, U_t, U_{t-9}, X\}_i$ for $i = 1, 2, ..., N$. The self-reported labor force status U_t is defined as follows:

$$U_t = \begin{cases} 1 & \text{employed} \\ 2 & \text{unemployed} \\ 3 & \text{not-in-labor-force} \end{cases}.$$

We denote the latent true labor force status at period t as U_t^*, which takes the same possible values as U_t. Let $\Pr(\cdot)$ stand for the probability distribution function of its argument, we outline our assumptions as follows.

Assumption 5.1.1 *The distribution of misclassification errors only depends the true labor force status in the current period, conditional on individual characteristics, i.e.,*

$$\Pr(U_t|U_t^*, X, \mathcal{U}_{\neq t}) = \Pr(U_t|U_t^*, X)$$

for all t with $\mathcal{U}_{\neq t} = \{(U_\tau, U_\tau^*),\ \text{for } \tau \neq t\}$.

Assumption 5.1.1 still allows the misclassification errors to be correlated with the true labor force status U_t^* and other variables in other periods through U_t^*. This is weaker than the classical measurement error assumption, where the error is independent of everything else, including the true values. Assumption 1 is a standard assumption in the literature and allows the misreporting behavior to be summarized by a simple misclassification matrix. Moreover, Meyer (1988) examines this assumption and finds it likely to be valid for CPS data. Assumption 5.1.1 implies that the joint probability of the observed labor force status $\Pr(U_{t+1}, U_t, U_{t-9}|X)$ is associated with the unobserved ones as follows:

$$\begin{aligned} & \Pr(U_{t+1}, U_t, U_{t-9}|X) \qquad (5.3) \\ = & \sum_{U_{t+1}^*} \sum_{U_t^*} \sum_{U_{t-9}^*} \Pr(U_{t+1}|U_{t+1}^*, X) \Pr(U_t|U_t^*, X) \Pr(U_{t-9}|U_{t-9}^*, X) \Pr(U_{t+1}^*, U_t^*, U_{t-9}^*|X). \end{aligned}$$

Having established the conditional independence of the misclassification process, our next assumption deals with the dynamics of the latent true labor force status.

Assumption 5.1.2 *Conditional on individual characteristics, the true labor force status nine months ago has no predictive power over the true labor force status in the next period beyond the current true labor force status, i.e.,*

$$\Pr(U_{t+1}^*|U_t^*, U_{t-9}^*, X) = \Pr(U_{t+1}^*|U_t^*, X)$$

for all t.

close enough to minimize attrition in CPS samples; 2) we want the three months to cover the 8-month recess period in the CPS rotation structrue so that there are enough variations in the labor force status; 3) Assumption 2 on the dynamics of the latent true labor force status is more likely to be satisfied if we use the data reported a while ago, e.g., nine months earlier.

Biemer and Bushery (2000) impose a first-order Markov restriction on the dynamics of the latent labor force status, which states $\Pr\left(U_{t+1}^{*}|U_{t}^{*},U_{t-1}^{*},...,U_{1}^{*}\right)=\Pr\left(U_{t+1}^{*}|U_{t}^{*}\right)$. That assumption is likely to be too strong due to the presence of state dependency, serial correlation among idiosyncratic shocks, and unobserved heterogeneity (see e.g. Hyslop 1999). Our assumption 5.1.2 is considerably weaker because we use the true labor force status nine month ago. Under Assumption 5.1.2, equation (5.3) may be simplified as follows:

$$\begin{aligned}&\Pr\left(U_{t+1},U_{t},U_{t-9}|X\right)\\ =&\sum_{U_{t}^{*}}\Pr\left(U_{t+1}|U_{t}^{*},X\right)\Pr\left(U_{t}|U_{t}^{*},X\right)\Pr\left(U_{t}^{*},U_{t-9}|X\right).\end{aligned}\tag{5.4}$$

Following the identification results in Hu (2008), we show that all the probabilities containing the latent true labor force status U_t^* on the right-hand-side (RHS) of Equation (7.13) may be identified under reasonable assumptions. Integrating out U_{t+1} in Equation (7.13) leads to

$$\Pr\left(U_{t},U_{t-9}|X\right)=\sum_{U_{t}^{*}}\Pr\left(U_{t}|U_{t}^{*},X\right)\Pr\left(U_{t}^{*},U_{t-9}|X\right).\tag{5.5}$$

Following Hu (2008), we introduce our matrix notation. For any given subpopulation with individual characteristics $X=x$, we define the misclassification matrix as follows.

$$\begin{aligned}&M_{U_{t}|U_{t}^{*},x}\\ \equiv&\begin{bmatrix}\Pr\left(U_{t}=1|U_{t}^{*}=1,x\right) & \Pr\left(U_{t}=1|U_{t}^{*}=2,x\right) & \Pr\left(U_{t}=1|U_{t}^{*}=3,x\right)\\ \Pr\left(U_{t}=2|U_{t}^{*}=1,x\right) & \Pr\left(U_{t}=2|U_{t}^{*}=2,x\right) & \Pr\left(U_{t}=2|U_{t}^{*}=3,x\right)\\ \Pr\left(U_{t}=3|U_{t}^{*}=1,x\right) & \Pr\left(U_{t}=3|U_{t}^{*}=2,x\right) & \Pr\left(U_{t}=3|U_{t}^{*}=3,x\right)\end{bmatrix}\\ \equiv&\left[\Pr\left(U_{t}=i|U_{t}^{*}=k,X=x\right)\right]_{i,k}.\end{aligned}$$

Each column of the matrix $M_{U_t|U_t^*,x}$ describes how an individual (mis)reports his or her labor force status given a possible value of the true labor force status. The matrix $M_{U_t|U_t^*,x}$ contains the same information as the misclassification probabilities $\Pr\left(U_t|U_t^*,x\right)$, which means the identification of $M_{U_t|U_t^*,x}$ implies that of $\Pr\left(U_t|U_t^*,x\right)$. Similarly, we may define

$$\begin{aligned}M_{U_{t},U_{t-9}|x}&\equiv\left[\Pr\left(U_{t}=i,U_{t-9}=k|x\right)\right]_{i,k},\\ M_{U_{t}^{*},U_{t-9}|x}&\equiv\left[\Pr\left(U_{t}^{*}=i,U_{t-9}=k|x\right)\right]_{i,k},\\ M_{1,U_{t},U_{t-9}|x}&\equiv\left[\Pr\left(U_{t+1}=1,U_{t}=i,U_{t-9}=k|x\right)\right]_{i,k}.\end{aligned}$$

We also define a diagonal matrix as follows:

$$\begin{aligned}&D_{1|U_{t}^{*},x}\\ \equiv&\begin{bmatrix}\Pr\left(U_{t+1}=1|U_{t}^{*}=1,x\right) & 0 & 0\\ 0 & \Pr\left(U_{t+1}=1|U_{t}^{*}=2,x\right) & 0\\ 0 & 0 & \Pr\left(U_{t+1}=1|U_{t}^{*}=3,x\right)\end{bmatrix}.\end{aligned}$$

As shown in Hu (2008), Equations (7.13) and (5.5) imply the following two matrix

equations:

$$M_{1,U_t,U_{t-9}|x} = M_{U_t|U_t^*,x} D_{1|U_t^*,x} M_{U_t^*,U_{t-9}|x} \tag{5.6}$$

and

$$M_{U_t,U_{t-9}|x} = M_{U_t|U_t^*,x} M_{U_t^*,U_{t-9}|x}. \tag{5.7}$$

In order to solve for the unknown matrix $M_{U_t|U_t^*,x}$, we need a technical assumption as follows.

Assumption 5.1.3 *The distributions of the current self-reported labor force status conditional on different self-reported labor force statuses nine month ago are linearly independent, i.e.,* $\Pr(U_t|U_{t-9}=1,x)$ *is not equal to a linear combination of* $\Pr(U_t|U_{t-9}=2,x)$ *and* $\Pr(U_t|U_{t-9}=3,x)$ *for all* U_t *and* x.

This assumption is equivalent to the condition that the matrix $M_{U_t,U_{t-9}|x}$ is invertible. Since it is imposed directly on the observed probabilities, this assumption is directly testable. Under Assumption 5.1.3, Equation (5.7) implies that both $M_{U_t|U_t^*,x}$ and $M_{U_t^*,U_{t-9}|x}$ are invertible. Eliminating matrix $M_{U_t^*,U_{t-9}|x}$ in Equations (5.6) and (5.7) leads to

$$M_{1,U_t,U_{t-9}|x} M_{U_t,U_{t-9}|x}^{-1} = M_{U_t|U_t^*,x} D_{1|U_t^*,x} M_{U_t|U_t^*,x}^{-1}. \tag{5.8}$$

This equation implies that the observed matrix on the left-hand-side (LHS) has an eigenvalue-eigenvector decomposition on the RHS. The three eigenvalues are the three diagonal entries in $D_{1|U_t^*,x}$ and the three corresponding eigenvectors are the three columns in $M_{U_t|U_t^*,x}$. Note that each column of $M_{U_t|U_t^*,x}$ is a distribution so that the column sum is 1, which implies that the eigenvectors are normalized.

In order to make the eigenvector unique for each given eigenvalue, we need the eigenvalues to be distinctive, which is formally stated as follows.

Assumption 5.1.4 *A different true labor force status leads to a different probability of reporting "employed" in the next period, i.e.,* $\Pr(U_{t+1}=1|U_t^*=k,x)$ *are different for different* $k \in \{1,2,3\}$.

This assumption is also testable from Equation (5.8). This is because $\Pr(U_{t+1}=1|U_t^*=k,x)$ for $k \in \{1,2,3\}$ are eigenvalues of the observed matrix $M_{1,U_t,U_{t-9}|x} M_{U_t,U_{t-9}|x}^{-1}$. Therefore, Assumption 5.1.4 holds if and only if all the eigenvalues of $M_{1,U_t,U_{t-9}|x} M_{U_t,U_{t-9}|x}^{-1}$ in Equation (5.8) are distinct. Intuitively, this assumption implies that the true labor force status at period t has an impact on the probability of reporting to be employed one period later.

The distinct eigenvalues guarantee the uniqueness of the eigenvectors. Since we do not observe U_t^* in the sample, we need to reveal the value u_t^* for each eigenvector $\Pr(U_t|U_t^*=u_t^*,x)$. In other words, the ordering of the eigenvalues or the eigenvectors is still arbitrary in Equation (5.8). In order to eliminate this ambiguity, we make the following assumption.

Assumption 5.1.5 *Each individual is more likely to report the true labor force status than to report any other possible values, i.e.,*

$\Pr(U_t=k|U_t^*=k,x) > \Pr(U_t=j|U_t^*=k,x)$ *for* $j \neq k$.

This assumption does not reveal the value of these misclassification probabilities, nor require the probability of reporting the truth to be larger than 50%. Assumption 5.1.5 is consistent with results from CPS reinterviews (see e.g.: Poterba and Summers, 1984) and other validation studies discussed in Bound et al. (2001a).

Technically, Assumption 5.1.5 implies that the true labor force status is the mode of the conditional distribution of the self-reported labor force status in each column of the eigenvector matrix. Therefore, the ordering of the eigenvectors is fixed and the the eigenvector matrix $M_{U_t|U_t^*,x}$ is uniquely determined from the eigenvalue-eigenvector decomposition of the observed matrix $M_{1,U_t,U_{t-9}|x}M^{-1}_{U_t,U_{t-9}|x}$. In particular, after diagonalizing the directly-estimable matrix $M_{1,U_t,U_{t-9}|x}M^{-1}_{U_t,U_{t-9}|x}$, we rearrange the order of the eigenvectors such that the largest element of each column or each eigenvector, i.e, the mode of the corresponding distribution, is on the diagonal of the eigenvector matrix. Consequently, the misclassification probability $\Pr(U_t|U_t^*, X)$ may be expressed as a closed-form function of the observed probability $\Pr(U_{t+1}, U_t, U_{t-9}|X)$. Such a procedure is constructive because one may estimate the misclassification probability $\Pr(U_t|U_t^*, X)$ by following the identification procedure above.

We summarize the closed-form identification and estimation of the misclassification probability $\Pr(U_t|U_t^*, X)$ as follows.

Theorem 5.1.1 *Under Assumptions 5.1.1, 5.1.2, 5.1.3, 5.1.4, and 5.1.5, the misclassification matrix* $\Pr(U_t|U_t^*, X)$ *is uniquely determined by the observed joint probability of the self-reported labor force status in three periods, i.e.,* $\Pr(U_{t+1}, U_t, U_{t-9}|X)$, *through the unique eigenvalue-eigenvector decomposition in equation (5.8).*

Proof: The results directly follow from Theorem 1 in Hu (2008). A complete proof can be found in the online appendix.

Finally, we may estimate the distribution of the latent true labor force status $\Pr(U_t^*|X)$ using the misclassification probability $\Pr(U_t|U_t^*, X)$ from the following equation:

$$\Pr(U_t|X) = \sum_{U_t^*} \Pr(U_t|U_t^*, X) \Pr(U_t^*|X).$$

This equation implies

$$\begin{bmatrix} \Pr(U_t = 1|x) \\ \Pr(U_t = 2|x) \\ \Pr(U_t = 3|x) \end{bmatrix} = M_{U_t|U_t^*,x} \times \begin{bmatrix} \Pr(U_t^* = 1|x) \\ \Pr(U_t^* = 2|x) \\ \Pr(U_t^* = 3|x) \end{bmatrix}.$$

Since we have identified the misclassification probability $\Pr(U_t|U_t^*, X)$, we may solve for the distribution of the latent true labor force status $\Pr(U_t^*|X)$ from that of the self-reported labor force status $\Pr(U_t|X)$ by inverting the matrix $M_{U_t|U_t^*,x}$. Therefore, the distribution of the latent true labor force status for a given x is identified as follows:

$$\begin{bmatrix} \Pr(U_t^* = 1|x) \\ \Pr(U_t^* = 2|x) \\ \Pr(U_t^* = 3|x) \end{bmatrix} = M^{-1}_{U_t|U_t^*,x} \times \begin{bmatrix} \Pr(U_t = 1|x) \\ \Pr(U_t = 2|x) \\ \Pr(U_t = 3|x) \end{bmatrix}. \tag{5.9}$$

Given the marginal distribution of the demographic characteristics X, $\Pr(X)$, we may identify the marginal distribution of the latent true labor force status $\Pr(U_t^*)$ as follows

$$\Pr(U_t^*) = \sum_X \Pr(U_t^*|X) \Pr(X).$$

This gives the unemployment rate

$$\mu_t^* \equiv \frac{\Pr(U_t^* = 2)}{\Pr(U_t^* = 1) + \Pr(U_t^* = 2)},$$

and the labor force participation rate

$$\rho_t^* \equiv \Pr(U_t^* = 1) + \Pr(U_t^* = 2).$$

Our identification procedure is constructive as it leads directly to an estimator. A nice property of our approach is that if there is no misclassification error in the data, our estimator would produce the same unemployment rate and labor force participation rate as those based on the raw data, under the assumptions above. Our estimator does not require an initial consistent estimate or iterations as in the regular optimization algorithms do.

Evaluation of the Assumptions

Before proceeding to empirical work, we evaluate the key assumptions which are essential for our identification results. We perform extensive Monte Carlo simulations to examine the robustness of our estimator to deviations from Assumptions 1 and 2. We also test the validity of Assumptions 3 and 4 directly using CPS data. For Assumption 5, we argue that it is likely to hold based on previous empirical work in the literature. We summarize the main things we have done here while leaving all detailed results in the online appendix.

Assumption 1 imposes conditional independence of the misreporting process. We have considered three different kinds of deviations to this assumption in our Monte Carlo simulations. In the first case, we allow misreporting errors to be correlated with the latent true labor force status in the previous period, i.e., $\Pr(U_t|U_t^*, \mathcal{U}_{\neq t}) = \Pr(U_t|U_t^*, U_{t-1}^*)$. In the second case, misreporting errors may be correlated with the self-reported labor force status in the previous period, i.e., $\Pr(U_t|U_t^*, \mathcal{U}_{\neq t}) = \Pr(U_t|U_t^*, U_{t-1})$. Lastly, we consider a special case of a general relaxation of Assumption 1, i.e., $\Pr(U_t|U_t^*, \mathcal{U}_{\neq t}) = \Pr(U_t|U_t^*, U_{t-1}^*, U_{t-1})$, where people would report the same value as in the previous period with certain probability if their true labor force status does not change, otherwise, they would report following the baseline misclassification probability $\Pr(U_t|U_t^*)$.[7] In all cases, our simulation results show that our estimator is robust to reasonable deviations from Assumption 1.[8]

Similarly, Assumption 2 imposes conditional independence on the transition of the underlying true labor force status. In the Monte Carlo simulation setup, we relax this as-

[7] We do this in response to a referee's concern that reporting behaviors might be serially-correlated.

[8] The detailed Monte Carlo setup can be found at section 3.1.2 in the online appendix and the simulation results can be found at sections 3.2.2-3.2.4 in the online appendix.

sumption to allow the transition of the true LFS to depend on that 9 periods earlier, i.e., $\Pr(U^*_{t+1}|U^*_t, U^*_{t-9}) \neq \Pr(U^*_{t+1}|U^*_t)$. Our simulation results show that the estimator is robust to reasonable deviations to assumption 2.[9]

Assumption 3 requires an observed matrix to be invertible, and therefore, is directly testable from the CPS data. We use bootstrapping to show that the determinant of this matrix is significantly different from zero, which implies that the matrix is invertible.[10]

Under Assumptions 1, 2, and 3, Assumption 4 requires that the eigenvalues of an observed matrix be distinct. We may also directly test this assumption using the CPS data by estimating the differences between the eigenvalues. Our bootstrapping results show that the absolute differences between the eigenvalues are significantly different from zero, which implies that the eigenvalues are distinctive.[11]

Assumption 5 implies that each individual is more likely to report the true labor force status than any other possible values. We believe this assumption is intuitively reasonable. Also, we are not aware of any studies in the literature (see e.g. previous studies cited in our paper and those reviewed by Bound et al. (2001a)) that report anything in violation of this assumption.

5.1.3 Empirical Results

Matching of Monthly CPS Data

We use the public-use micro CPS data to estimate the unemployment rate and the labor force participation rate.[12] Each CPS monthly file contains eight rotation groups that differ in month-in-sample. The households in each rotation group are interviewed for four consecutive months after they enter, withdraw temporarily for eight months, then reenter for another four months of interviews before exiting the CPS permanently. Because of the rotational group structure, the CPS can be matched to form longitudinal panels, which enable us to obtain the joint probabilities of the self-reported labor force statuses in three periods.

We follow the algorithm proposed by Madrian and Lefgren (2000) to match adjacent CPS monthly files.[13] There are two main steps in the process of matching. First, the CPS samples are matched based on identifiers. If two individuals in two CPS monthly files (within the corresponding rotational groups) have the same household identifier, household replacement number (which denotes whether this is a replacement of the initial household) and personal identifier (which uniquely identifiers a person within a household), then the two individuals are declared as a "crude match". This step is not perfect and may result in considerable matching errors because there might exist coding errors with respect to those

[9] The detailed Monte Carlo setup can be found at section 3.1.3 in the online appendix and the simulation results can be found at section 3.2.5 in the online appendix.

[10] Detailed results can be found at section 4 (Table A11) in the online appendix.

[11] Results can be found at Table A12 of section 4 in the online appendix.

[12] All data are downloaded from www.bls.census.gov/cps_ftp.html. Following BLS practices, we restrict the samples to those aged 16 and over. Sample summary statistics can be found at Table A1 in the online appendix.

[13] See also Feng (2001) and Feng (2008).

identifiers. Therefore, the second step uses information on sex, age and race to "certify" the crude match. In the matching algorithm we use, if the sex or race reported in the two monthly files corresponding to a crude match are different, or if the age difference is greater than 1 or less than 0, then we discard the match as a false one.

As the previous literature (e.g.: Peracchi and Welch 1995 and Feng 2008) has documented, the matched sample is not representative of the cross-sectional sample in period t due to sample attrition in matching. We use the matching weights to correct for attrition. First, we run a Logit regression for the period t cross-sectional sample, where the dependent variable is either 1 (the observation is matched) or 0 (the observation is not matched), and the independent variables are sex, race, age, schooling, and the labor force status in period t. We next calculate the predicted probabilities of being matched for all the observations in the matched sample. The final matched sample is then weighted using the inverse of the predicted match probabilities. This adjustment procedure ensures the cross-sectional sample and the matched sample have the same marginal distributions on the key individual characteristics for period t.[14]

Misclassification Probabilities

For each demographic group, we pool matched samples to estimate the misclassification probabilities.[15] Table 1 reports results for all the eight groups, including (1) white males aged 40 and younger; (2) white males aged over 40; (3) nonwhite males aged 40 and younger; (4) nonwhite males aged over 40; (5) white females aged 40 and younger; (6) white females aged over 40; (7) nonwhite females aged 40 and younger; (8) nonwhite females aged over 40. There exist some consistent patterns across all the groups. When the actual labor force status is either employed or not-in-labor-force, the probabilities of being misreported to a different labor force status are typically small and never above 6%. The biggest errors come from the unemployed people being misclassified as either not-in-labor-force or employed. Only around 50-70% of unemployed people correctly report their true labor force status. For example, for white males aged 40 and younger, 20% of the unemployed report to be employed, while another 17% of them report as not-in-labor-force. On the other hand, there are considerable heterogeneities among different demographic groups. For example, 10.8% of the unemployed white females aged 40 and younger report as not-in-labor-force, while all other groups have much higher probabilities of reporting to be not-in-labor-force while unemployed.

[14]Under the assumption that attrition is solely based on observables, our correction method using match weights is consistent. To check for robustness of our procedure we have also tried not using matching weights, i.e., not correcting for attrition in matching, and found similar results in terms of corrected unemployment rates. Details can be found at section 5.5 of the online appendix.

[15]To be consistent with the last version of the paper we pool data from January 1996 to December 2009. The estimated misclassification probabilities do not change statistically significantly if we pool all data up to August 2011. Please refer to section 5.3 of the online appendix for details and more elaborate discussions.

Table 1: Misclassification probabilities (%) for different demographic groups

	$\Pr(i\|j) \equiv \Pr(U_t = i\|U_t^* = j)$					
Demographic group	Pr(2|1)	Pr(3|1)	Pr(1|2)	Pr(3|2)	Pr(1|3)	Pr(2|3)
(1) Male/White/age≤40	0.9	1.3	20.1	17.2	6.0	0.0
	(0.06)	(0.07)	(1.28)	(2.69)	(0.42)	(0.39)
(2) Male/White/age > 40	0.4	0.9	16.5	18.8	1.4	0.1
	(0.03)	(0.05)	(1.14)	(2.34)	(0.07)	(0.07)
(3) Male/Nonwhite/age ≤ 40	1.1	2.2	13.4	18.1	5.0	4.3
	(0.10)	(0.13)	(1.21)	(3.91)	(0.36)	(1.26)
(4) Male/Nonwhite/age > 40	0.7	1.5	15.5	22.0	1.2	0.0
	(0.08)	(0.10)	(1.81)	(5.55)	(0.16)	(0.12)
(5) Female/White/age≤40	0.6	2.1	18.6	10.8	4.4	0.0
	(0.05)	(0.10)	(1.59)	(4.10)	(0.27)	(0.08)
(6) Female/White/age>40	0.3	1.4	17.9	28.2	1.0	0.0
	(0.03)	(0.07)	(1.46)	(3.16)	(0.06)	(0.01)
(7) Female/Nonwhite/age≤40	1.1	2.6	11.8	29.4	2.2	0.0
	(0.09)	(0.16)	(1.54)	(8.24)	(0.70)	(0.01)
(8) Female/Nonwhite/age>40	0.4	1.8	13.9	25.0	1.2	0.7
	(0.07)	(0.11)	(1.89)	(5.82)	(0.09)	(0.17)
Overall	0.6	1.5	17.3	20.2	2.9	0.2
	(0.02)	(0.03)	(0.59)	(1.39)	(0.10)	(0.09)

Note: Bootstrap standard errors based on 500 repetitions are reported in parentheses.

We also formally test for the differences in the misclassification probabilities between the groups. For example, we consider males vs. females, controlling for race and age categories. We find that employed males are more likely to misreport as unemployed but less likely to misreport as not-in-labor-force than employed females. The differences are always statistically significant at the 5% significance level except for the comparison between nonwhite males aged 40 and younger and nonwhite females aged 40 and younger. When unemployed, the differences are mostly insignificant, with the only exception being that white males aged over 40 are less likely to misreport as being not-in-labor-force compared to white females aged over 40. In addition, when not-in-labor-force, males are more likely to be misclassified as employed.[16]

Some previous studies have made strong assumptions regarding between-group misclassification errors. For example, in order to achieve identification, Sinclair and Gastwirth (1998) assume that males and females have the same misclassification error probabilities (see also Sinclair and Gastwirth 1996), which we can safely reject.[17] In general, our results suggest that the equality assumptions of misclassification probabilities across different demographic groups, which are essential for identification in the H-W models, are unlikely to hold in reality.

The last two rows of Table 1 report misclassification probabilities and associated standard errors for the overall U.S. population. The results are broadly consistent with those in the existing literature. When we compare our estimates of misclassification probabilities with some of those obtained in the existing literature,[18] we see the same general pat-

[16]Comparisons between males and females and other demographic characteristics can be found in Table A13 in the online appendix.

[17]See the first panel in Table A13 in the online appendix.

[18]These estimates can be found in Table A14 in the online appendix.

tern: the biggest misclassification probabilities happen when unemployed individuals misreport their labor force statuses as either not-in-labor-force ($\Pr(U_t = 3|U_t^* = 2)$) or employed ($\Pr(U_t = 1|U_t^* = 2)$), while the other misclassification probabilities are all small. Our point estimates of $\Pr(U_t = 3|U_t^* = 2)$ and $\Pr(U_t = 1|U_t^* = 2)$ are somewhat higher than many of the existing estimates. But our estimates are well within the 95% confidence intervals in many existing studies because of their large standard errors. Due to our methodological advantages and the large sample size we use, we are able to produce much more precise estimates.

The Unemployment Rate

Given the estimated misclassification matrices, we then calculate distribution of the latent true labor force status for each demographic group based on Equation (5.9). To estimate $\Pr(U_t|X)$, we use all the eight rotation groups in any given CPS monthly file, which subsequently give us the self-reported unemployment rate and the labor force participation rate. Once we have $\Pr(U_t^*|X)$, we can calculate the corrected unemployment rate and the corrected labor force participation rate. In order to be consistent with officially-announced statistics, all numbers are weighted using final weights provided by CPS.[19]

Table 2 presents the results for each demographic group. We divide the study period into three sub-periods based on the US business cycles.[20] The first sub-period goes from January 1996 to October 2001, which is roughly the end of the 2001 recession. The second sub-period is from November 2001 to November 2007, corresponding to the expansion period between two recessions (the 2001 recession and the most recent 2007-09 recession). The third sub-period goes from December 2007 to the end of our study period, i.e., Aug 2011, which includes the 2007-2009 recession and its aftermath.

For each demographic group and each sub-period, the corrected unemployment rates are always higher than the reported ones. Note also that for all demographic groups, sub-period 3 posts the highest levels of unemployment, followed by sub-period 2, and then by the first sub-period. This relationship is unchanged using either the reported or the corrected rates. In addition, the degree of underestimation is larger when the level of unemployment is higher. For example, for white males less than 40, in the first sub-period, the corrected unemployment rate is 6.5%, which is higher than the reported unemployment rate by 1.5 percentage points. In the second sub-period, the corrected unemployment rate is 8.2%, which is higher than the reported unemployment rate by 2.1%. The largest differential appears in the latest recession period. In this case, the corrected unemployment rate is 14.5%, which is higher than the reported unemployment rate by 4.4% – a 44% upward adjustment.

Table 2: Unemployment rates (%) averaged over three sub-periods for different demographic groups

[19]The final weights in the CPS micro data have been adjusted for a composite estimation procedure that BLS uses to produce official labor force statistics (Appendix I in BLS, 2000).

[20]see http://www.nber.org/cycles.html.

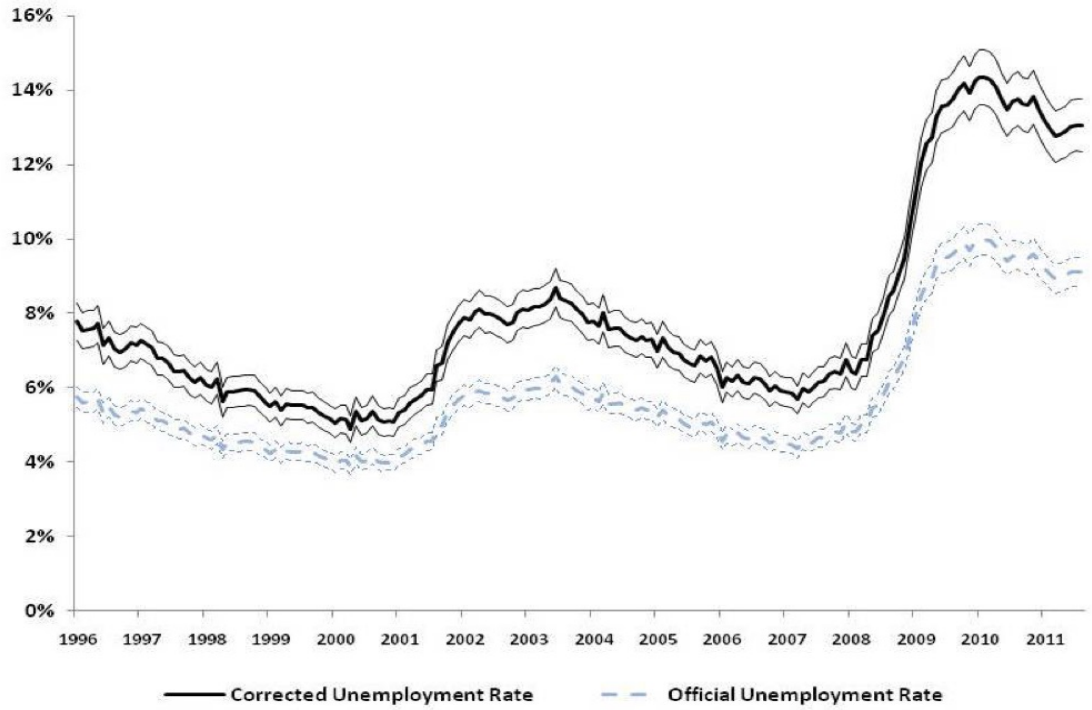
Demographic group	Sub-period 1 (1996/01-2001/10)		Sub-period 2 (2001/11-2007/11)		Sub-period 3 (2007/12-2011/8)	
	reported	corrected	reported	corrected	reported	corrected
(1) Male/White/age≤40	5.0	6.5	6.1	8.2	10.1	14.5
	(0.2)	(0.4)	(0.3)	(0.5)	(0.5)	(0.8)
(2) Male/White/age>40	2.7	3.4	3.4	4.5	6.3	8.9
	(0.1)	(0.2)	(0.2)	(0.2)	(0.3)	(0.5)
(3) Male/Nonwhite/age≤40	10.1	11.1	10.8	12.0	16.0	19.3
	(0.5)	(0.9)	(0.5)	(1.1)	(0.7)	(1.4)
(4) Male/Nonwhite/age>40	4.8	6.5	5.8	8.0	9.6	13.9
	(0.2)	(0.5)	(0.3)	(0.6)	(0.4)	(1.0)
(5) Female/White/age≤40	5.1	6.4	5.8	7.3	8.3	10.9
	(0.2)	(0.4)	(0.3)	(0.5)	(0.4)	(0.7)
(6) Female/White/age>40	2.7	4.4	3.2	5.3	5.4	9.1
	(0.1)	(0.3)	(0.1)	(0.3)	(0.2)	(0.5)
(7) Female/Nonwhite/age≤40	10.0	14.5	10.3	14.9	13.4	19.8
	(0.5)	(1.5)	(0.5)	(1.6)	(0.6)	(2.0)
(8) Female/Nonwhite/age>40	4.2	5.1	5.2	6.8	7.2	10.0
	(0.2)	(0.5)	(0.2)	(0.6)	(0.3)	(0.9)
Overall	4.4	5.9	5.1	6.9	8.1	11.5
	(0.1)	(0.2)	(0.1)	(0.2)	(0.1)	(0.3)

Note: Numbers reported in parentheses are bootstrap standard errors based on 500 repetitions.

We then estimate the unemployment rates and the corresponding standard errors for the US population as a whole, based on the results for all the demographic groups. Based on the last two rows of Table 2, corrected unemployment rates for the US population are 5.9%, 6.9% and 11.5% for the three sub-periods, respectively. Note that the degree of underestimation is substantially larger in the third sub-period, official unemployment rate is 3.4 percentage points lower than the corrected one, while in the first two sub-periods the discrepancies are only 1.5 and 1.8 percentage points, respectively. Figure 1 displays all the monthly values that are seasonally-adjusted. For the whole period, the corrected unemployment rate is always higher than the reported one and the difference is between 1% and 4.4%, and 2.1% on average.

The substantial degree of underestimation of the unemployment rate may not be very surprising because most of the misclassification errors are from the unemployed people misreporting their labor force status as either employed or not-in-labor-force. We believe this arises primarily due to the intrinsic difficulties in classifying labor force status for some specific groups of people. Among those not-in-labor-force, marginally-attached workers, especially discouraged workers, could be classified as unemployed because they also desire a job although do not search in the job market. In fact, Jones and Riddell (1999) find that some marginally-attached workers are behaviorally more similar to unemployed than to the rest of those not-in-labor-force. On the other hand, involuntary part-time workers are classified as employed according to the official definition. But many of them could be

Figure 1: Corrected and official (reported) unemployment rates



Note: Figure displays seasonally-adjusted corrected unemployment rates (in solid line) and official unemployment rates (in dashed line) for the whole population from Jan 1996 to Aug 2011. The corresponding thin lines signify 95% upper and lower confidence bounds. For seasonally adjustment, we use Census Bureau's WinX12 software.

observationally more similar to unemployed workers.[21] [22]

Table 3 decomposes the underestimation of unemployment rate. For the period of January 1996 to August 2011, the official statistics underestimate the unemployment rate on average by 2.1 percentage points. The degree of underestimation varies, however, by demographic group. On the one hand, the young nonwhite female group posts the largest level of underestimation, at 5 percentage points. On the other hand, the official statistics only underestimate by 1.3 percentage points for white males over 40. In terms of contributions to the total degree of underestimation (last column of Table 3), white females over 40 declare the largest share of the total (27%), followed by white males 40 and younger (21%). Nonwhite groups contributed relatively little as they account for relatively small portions of the US total population.

[21] For example, Farber (1999) examine displaced workers and find temporary and involuntary part-time jobs are part of the transitional process from unemployment to full-time work.

[22] According to the broadest concept of unemployment by BLS, U6, all marginally-attached workers and involuntary part-time workers are counted as unemployed. Our corrected unemployment rate series are substantially lower than U6, as shown by Figure A4 in the online appendix.

Table 3: Decomposition of underestimation in unemployment rates by demographic groups

Demographic group	Underestimation in unemployment rate $(a) = \widehat{\mu_t^*} - \mu_t$	Group share in US population (b)	Contribution to underestimation $(c) = (a) \times (b)$	Relative contribution $(d) = \frac{(c)}{\sum(c)}$
(1) Male/White/age≤40	2.41	18.24	0.44	20.57
(2) Male/White/age>40	1.34	21.80	0.29	13.65
(3) Male/Nonwhite/age≤40	1.72	4.46	0.08	3.59
(4) Male/Nonwhite/age>40	2.65	3.73	0.10	4.63
(5) Female/White/age≤40	1.68	17.91	0.30	14.08
(6) Female/White/age>40	2.37	24.20	0.57	26.82
(7) Female/Nonwhite/age≤40	5.05	4.99	0.25	11.79
(8) Female/Nonwhite/age>40	1.76	4.68	0.08	3.86
Total		100.00	2.14	100.00

Note: Table reports averages over the January 1996 to August 2011 period. All numbers are rounded. (a) Underestimation in the unemployment rate (%), which equals the average corrected unemployment rate $\widehat{\mu_t^*}$ minus the average official unemployment rate μ_t; (b) Population share of the demographic group; (c) Contribution to the total US underestimation in the unemployment rate (%), which equals (a) times (b); (d) Relative contribution to the total underestimation, which equals (c) divided by its column sum.

One particular concern is whether misclassification behaviors and the resulted corrected unemployment rates would depend on labor market conditions. For example, when the labor market is weak and the pool of unemployed people includes a larger share of job losers and others whose status is unambiguous, then the misreporting of unemployment would tend to be less prevalent. In order to test this hypothesis directly, we have estimated three different misclassification probabilities for each demographic group for the three sub-periods. We do find some evidence that the misclassification probabilities are different in different sub-period corresponding to different labor market conditions. More specifically, sub-period 3 (December 2007 to August 2011), which is characterized by much higher rate of unemployment and presumably much weaker labor market conditions compared to the previous two sub-periods, has lower levels of misclassification in general. Nevertheless, we show that the corrected unemployment series are robust to whether we allow misclassification probabilities to differ in different sub-periods.[23].

We have also examined the effect of misclassification on the labor force participation (LFP) rates. For each demographic group for the three sub-periods: January 1996 to October 2001, November 2001 to November 2007, and December 2007 to August 2011, the corrected labor force participation rates are always higher than the reported ones, but the differences are small and not statistically significant. For the US population as a whole, average difference between corrected and official LFP rates is less than 2%, and not statistically significant. For the three sub-periods, the corrected labor force participation rates are 68.1%, 67.3% and 66.8%, respectively. The reported rates are only slightly lower, at 67.1%, 66.2% and 65.2%, respectively. [24] Therefore, misclassification errors cause little change to the labor force participation rate. Compared with the number of unemployed

[23]Detailed results can be found at section 5.4 in the online appendix.

[24]Detailed results are shown in section 7 of the online appendix.

people, the total number of people who are in labor force is much larger. Hence any corrections due to misclassification errors will have a relatively small effect.

5.1.4 Summary

This paper examines misclassification errors in labor force status using CPS data. Similar to previous studies, we show that there exist considerable misclassifications from unemployed to not-in-labor-force and from unemployed to employed. The results at least partly reflect the intrinsic difficulties in classifying labor force statuses of certain groups of people, such as marginally attached workers (especially discouraged workers) and part-time workers for economic reasons, into three distinct categories. We correct for such errors and show that the official U.S. unemployment rate significantly underestimates the true level of unemployment in the United States. For the period from January 1996 to August 2011, our corrected unemployment rates are higher than the reported ones by 2.1 percentage points on average, with differences ranging from 1 to 4.4 percentage points and always statistically significant. In addition, our estimates suggest that unemployment might be much more sensitive to business cycles than previously thought, as the degree of underestimation is larger in magnitude when unemployment rate is higher.

5.2 Cognitive and Noncognitive Skill Formation

Cunha et al. (2010) consider a model of cognitive and non-cognitive skill formation, where for multiple periods of childhood $t \in \{1, 2, \ldots, T\}$, $X_t^* = \left(X_{C,t}^*, X_{N,t}^*\right)$ stands for cognitive and non-cognitive skill stocks in period t, respectively. The T childhood periods are divided into $s \in \{1, 2, \ldots, S\}$ stages of childhood development with $S \leq T$. Let $I_t = (I_{C,t}, I_{N,t})$ be parental investments at age t in cognitive and non-cognitive skills, respectively. For $k \in \{C, N\}$, they assume that skills evolve as follows:

$$X_{k,t+1}^* = f_{k,s}\left(X_t^*, I_t, X_P^*, \eta_{k,t}\right), \tag{5.10}$$

where $X_P^* = \left(X_{C,P}^*, X_{N,P}^*\right)$ are parental cognitive and non-cognitive skills and $\eta_t = (\eta_{C,t}, \eta_{N,t})$ is random shocks. If one observes the joint distribution of X^* defined as

$$X^* = \left(\left\{X_{C,t}^*\right\}_{t=1}^T, \left\{X_{N,t}^*\right\}_{t=1}^T, \{I_{C,t}\}_{t=1}^T, \{I_{N,t}\}_{t=1}^T, X_{C,P}^*, X_{N,P}^*\right), \tag{5.11}$$

one can estimate the skill production function $f_{k,s}$.

However, the vector of latent factors X^* is not directly observed in the sample. Instead, they use measurements of these factors satisfying

$$X_j = g_j\left(X^*, \varepsilon_j\right) \tag{5.12}$$

for $j = 1, 2, \ldots, M$ with $M \geq 3$. The variables X_j and ε_j are assumed to have the same

dimension as X^*. Under the assumption that

$$X_1 \perp X_2 \perp X_3 \mid X^*, \tag{5.13}$$

this leads to a 3-measurement model and the distribution of X^* can then be identified from the joint distribution of the three observed measurements. The measurements X_j in their application include test scores, parental and teacher assessments of skills, and measurements on investment and parental endowments. While estimating the empirical model, they assume a linear function g_j and use Kotlarski's identity to directly estimate the latent distribution.

5.3 Income dynamics

The literature on income dynamics has been focusing mostly on linear models, where identification is usually not a major concern. When income dynamics have a nonlinear transmission of shocks, however, it is not clear how much of the model can be identified. Arellano et al. (2017) investigate the nonlinear aspect of income dynamics and also assess the impact of nonlinear income shocks on household consumption.

They assume that the pre-tax labor income Y_{it} of household i at age t satisfies

$$Y_{it} = \eta_{it} + \varepsilon_{it} \tag{5.14}$$

where η_{it} is the persistent component of income and ε_{it} is the transitory one. Furthermore, they assume that ε_{it} has a zero mean and is independent over time, and that the persistent component η_{it} follows a first-order Markov process satisfying

$$\eta_{it} = Q_t\left(\eta_{i,t-1}, u_{it}\right) \tag{5.15}$$

where Q_t is the conditional quantile function and u_{it} is uniformly distributed and independent of $(\eta_{i,t-1}, \eta_{i,t-2}, \ldots)$. Such a specification is without loss of generality under the assumption that the conditional CDF $F\left(\eta_{it}|\eta_{i,t-1}\right)$ is invertible with respect to η_{it}.

The dynamic process $\{Y_{it}, \eta_{it}\}$ can be considered as a hidden Markov process as $\{X_t, X_t^*\}$ in equations (2.73) and (2.74). As we discussed before, the nonparametric identification is feasible with three periods of observed income $(Y_{i,t-1}, Y_{it}, Y_{i,t+1})$ satisfying

$$Y_{i,t-1} \perp Y_{it} \perp Y_{i,t+1} \mid \eta_{it} \tag{5.16}$$

which forms a 3-measurement model. Under the assumptions in Theorem 2.4.2, the distribution of ε_{it} is identified from $f\left(Y_{it}|\eta_{it}\right)$ for $t = 2, \ldots, T-1$. The joint distribution of η_{it} for all $t = 2, \ldots, T-1$ can then be identified from the joint distribution of Y_{it} for all $t = 2, \ldots, T-1$. This leads to the identification of the conditional quantile function Q_t.

For a non-Markovian process, Hu et al. (2018) consider the canonical model of earnings dynamics developed in the 1970s and 1980s, which includes a random walk permanent component and an ARMA transitory component, with the un- derlying permanent and

transitory unobservable shocks assumed to be i.i.d. but otherwise unspecified. The observed earnings Y_t in year t is decomposed into two independent components:

$$Y_t = U_t + V_t. \tag{5.17}$$

The first one, U_t is the permanent component which follows the unit root process:

$$U_t = U_{t-1} + u_t, \tag{5.18}$$

where u_t is the permanent shock. The second one, V_t is the transitory component which follows the ARMA(p,q) process:

$$V_t = \rho_{t,1} V_{t-1} + \rho_{t,2} V_{t-2} \cdots + \rho_{t,p} V_{t-p} + G_t\left(\epsilon_t, \epsilon_{t-1}, \cdots, \epsilon_{t-q}\right), \tag{5.19}$$

Define $Y_{t+1} = Y_{t+1} - Y_t$. For an ARMA(1,1) process, they show that the AR coefficient can be directly estimated up to a normalization as follows:

$$\rho_{t+1} \frac{1-\rho_{t+2}}{1-\rho_{t+1}} = \frac{cov\left(\Delta Y_{t+2}, Y_{t-1}\right)}{cov\left(\Delta Y_{t+1}, Y_{t-1}\right)}$$

Furthermore, they show

$$\begin{aligned} Y_t &= V_t + U_t \\ \frac{\Delta Y_{t+2}}{\rho_{t+2}-1} - \Delta Y_{t+1} &= V_t + \frac{G_{t+2}(\epsilon_{t+2}, \epsilon_{t+1}) + u_{t+2}}{\rho_{t+2}-1} - u_{t+1} \end{aligned}$$

The Kotlarski's identity then implies that the distribution of V_t can be identified with a closed-form. In the end, they show that the joint distribution of $\{Y_t\}_{t=1,\ldots,T\geqslant 3}$ uniquely determines distributions of latent variables u_t, ϵ_t, U_t, and V_t. Although this model imposes parametric restrictions, such as random work and ARMA structures, the distributions of shocks are left nonparametrically. The identification of such semiparametric dynamic models with latent variables is complimentary to the existing results, which all heavily rely on a Markovian property of the dynamic structure. Hu et al. (2018) is the first to show the identification of a non-Markovian process with latent variables. Their identification results open the possibility of identification of more general non-Markovian processes with latent variables, which could have broad applications in empirical research. We provide details in Hu et al. (2018) as follows.

5.3.1 Background

Methods of estimating models with panel data have a long history. Those methods were first developed in the 1950s and 1960s for panel data sets of firms and of state aggregates for consumption (see Nerlove (2002) for a recounting of this period of development and for the key historical references). What we term the "canonical" model was developed in that period, consisting of a permanent component and a transitory component, distributed independently of each other. In some variants, the transitory component was assumed to

follow a simple low-order ARMA process. Because of its simplicity, its intuition, and its alignment with economic theories which have permanent and transitory processes, the model has been enormously influential and has found applications in dozens of areas. Models of earnings dynamics, consumption dynamics, dynamics for firms or industries, and dynamics for individual health, student academic achievement, and other individual outcomes are just a few examples of applications.

This paper considers the identification and estimation of the canonical model under nonparametric assumptions on the unobservables. While the literature on panel data models since their development is enormous, most papers have generalized the model with additional parametric features (random walks, random growth terms, higher-order ARMAs, and other stochastic processes) and most have concerned themselves with fitting the parameters of the model only to the second moments of the data and hence fitting only the second moments of the unobservables. Our goals are to determine under what assumptions the full distribution of the unobservables in the model can be nonparametrically identified, to provide an estimator for the relevant distributions, and to provide an empirical application.

We first establish identification for our model, which is a somewhat modified version of the canonical model in several respects. For example, we allow a slightly generalized version of the common MA process, allowing it to be nonlinear; we allow the AR process to be nonstationary and to change with age; and we do not assume the shocks in each period to be i.i.d. We prove identification of the model by showing that the key unobserved elements have repeated measurements with classical measurement errors. We can, therefore, make use of the Kotlarski's identity (Kotlarski, 1967; Rao, 1992; Li and Vuong, 1998; Schennach, 2004; Bonhomme and Robin, 2010; Evdokimov, 2010) to provide closed-form identification of the distribution of the unobservables. In the identification of the generalized MA process, we rely on a recently developed result for nonlinear measurement error models (Schennach and Hu, 2013). We also provide an estimator based on deconvolution methods, which is similar to the existing estimators developed for this closed-form identification results (Li and Vuong, 1998). An advantage of this closed-form estimator is that it requires many fewer nuisance parameters than alternative semiparametric estimators.

Prior work on nonparametric identification and estimation of the canonical model and expanded versions of it include Horowitz and Markatou (1996) and Bonhomme and Robin (2010). Our paper differs from those by its approach. While the existing identification results for dynamic models with latent variables rely on a Markovian property of the dynamic structure, our paper complements the existing literature by showing the identification of a semiparametric unit-root process of a permanent state variable and a semiparametric non-Markovian process of a transitory state variable. In particular, the transitory state variable is generated by an ARMA process and does not follow a finite-order Markov process.[25] Nonparametric approaches applied to earnings dynamics models have also been developed by Geweke and Keane (2000), who allow some of the unobservables to be a mixture of normals, and by Arellano, Blundell, and Bonhomme (2017), who replace the unit root process

[25]In fact, the AR process is a higher-order Markov process, but the MA process is not a finite-order Markov.

on the permanent component with a nonparametric autoregressive function while maintaining an independence assumption for the transitory error. Our model keeps the unit root process and allow the transitory shocks to follow a semiparametric ARMA process as in the canonical models. As mentioned above, such a process of the transitory state is not Markovian and therefore can capture different dynamic structures. As for methodology, Arellano, Blundell, and Bonhomme (2017) use the results in Hu and Schennach (2008) for a general nonlinear nonclassical measurement error model with three observables. Our paper uses the Kotlarski's identity (Kotlarski, 1967; Rao, 1992; Li and Vuong, 1998; Schennach, 2004; Bonhomme and Robin, 2010; Evdokimov, 2010) and the results in Schennach and Hu (2013) for a nonlinear model with classical measurement errors when only two observables are available.

We also provide an application to the earnings dynamics of U.S. men using the Panel Study on Income Dynamics (PSID), the data set most commonly used in the literature on estimating models of individual earnings dynamics. There is a very large literature on applications to earnings dynamics models, going back to early work by Hause (1977), Lillard and Willis (1978), MaCurdy (1982), and Abowd and Card (1989), followed by many contributions including those by Horowitz and Markatou (1996), Baker (1997), Meghir and Pistaferri (2004), Guvenen (2007, 2009), Bonhomme and Robin (2010), Browning, Ejrnaes, and Alvarez (2010), Hryshko (2012), Jensen and Shore (2014), Arellano, Blundell, and Bonhomme (2017), and Botosaru and Sasaki (2018). A review of this literature, including studies which have allowed the dynamic processes to shift with calendar time, can be found in Moffitt and Zhang (2018).

Our results show that the marginal distributions of log earnings of U.S. men are non-normal, with significant skewness and fatter tails of both the permanent and transitory components of earnings than the normal. We also find earnings dynamics very different than the normal, for our results show that the likelihood of remaining in a lower tail of the permanent earnings distribution does not fall over time as much, suggesting considerably less earnings mobility than would be found with a multivariate normality assumption. Another important finding from our empirical analysis is that the estimates of the marginal distributions as well of persistence and dynamics of permanent earnings are very sensitive to the degree of persistence in the transitory component. We find evidence for the existence of higher-order ARMA processes in the transitory component and that, with such higher-order processes, the permanent component of earnings has much less variability in marginal distributions and less mobility over time. Thus the transitory component makes a much stronger relative contribution to the marginal earnings distributions and to earnings mobility than in much of the prior literature, which often allows much less persistence in the transitory component. Finally, we consider earnings dynamics in subsamples of men with strong labor force attachment and of married men (both subsamples have been studied in the literature), finding both subsamples to have lower variances of permanent and transitory shocks than for the full population but also more earnings mobility than that population.

5.3.2 A Semiparametric Canonical Permanent-Transitory Model

We consider the following setup of a semiparametric state space model. The measurement Y_t in time t is decomposed into two independent components:

$$Y_t = U_t + V_t. \tag{5.20}$$

The first one, U_t is the permanent state which follows the unit root process:

$$U_t = U_{t-1} + \eta_t \tag{5.21}$$

with innovation η_t. The second one, V_t is the transitory state which follows the ARMA(p,q) process:

$$V_t = \rho_{t,1} V_{t-1} + \rho_{t,2} V_{t-2} \cdots + \rho_{t,p} V_{t-p} + G_t\left(\varepsilon_t, \varepsilon_{t-1}, \cdots, \varepsilon_{t-q}\right). \tag{5.22}$$

For a short-hand notation, we write the vector of the AR coefficients by $\rho_t = (\rho_{t,1}, \cdots, \rho_{t,p})'$. Note that the time effect is the source of non-stationarity in this model both through the time-varying ARMA specifications (i.e, ρ_t and G_t) and through arbitrary time variations in the distributions of the primitives (i.e., η_t and ε_t). Because of the nonparametric specification of these time-varying distributions of the primitives, the time effect may appear in higher-order moments as well as in the first moment e.g., as commonly introduced by additive time effects in (5.20), as is common in applications. In contrast to much of the literature, we allow arbitrarily high-order ARMA processes and this will be a major feature of our empirical application in Section 6.

Our first goal in this paper is the identification of the nonparametric distributions of U_t, V_t, η_t, and ε_t as well as the function G_t and the AR parameters ρ_t in this state space model. The following example illustrates an application of this general framework to a semiparametric model of earnings dynamics.

Example 1 (The Model of Earnings Dynamics) *One application is the model of earnings dynamics, where the measurement Y_t is the observed earnings at age t, the permanent state U_t is the permanent component of earnings at age t, the innovation η_t is the permanent shock at age t, and the transitory state V_t is the transitory component of earnings at age t.*

5.3.3 An Illustration of the Identification Strategy

For an illustration, we focus on the model where the permanent state follows the unit root process and the transitory state follows an ARMA(1,1) process. The general identification results will follow in Section 6.2.3. In a random sample, we observe the joint distribution of Y_t for periods $t = 1, 2, \cdots, T$. While we keep the parts (5.20) and (5.21) of the general model, the ARMA part (5.22) simplifies to

$$V_t = \rho_t V_{t-1} + G_t\left(\varepsilon_t, \varepsilon_{t-1}\right) \tag{5.23}$$

in the current section. The unknown coefficient ρ_t and the unknown function G_t may be time-varying. Furthermore, we do not require a parametric or semiparametric specification

of G_t. We assume the following independence condition.

Assumption 5.3.1 *(i) The random variables $\eta_T, \cdots, \eta_1$, U_0, $\varepsilon_T, \cdots, \varepsilon_1$, and the random vector (ε_0, V_0) are mutually independent, i.e.,*

$$f(\eta_T, \cdots, \eta_1, U_0, \varepsilon_T, \cdots, \varepsilon_1, \varepsilon_0, V_0) = f(\eta_T) \cdots f(\eta_1) f(U_0) f(\varepsilon_T) \cdots f(\varepsilon_1) f(\varepsilon_0, V_0).$$

(ii) $(\eta_T, \cdots, \eta_1, U_0, V_0)$ have zero means and $E[G_t(\varepsilon_t, \varepsilon_{t-1})] = 0$ for $t \in \{1, \cdots, T\}$.

This assumption implies that process $\{U_t\}$ is independent of process $\{V_t\}$. We leave the marginal distributions of η_t and ε_t unspecified and allow them to vary arbitrarily with t. In this setup, we are interested in identification of the nonparametric distributions of the primitives ε_t and η_t, the structures ρ_t and G_t, and the nonparametric distributions of the components U_t and V_t. Our identification strategy is illustrated below in four steps.

Step 1: Identification of f_{V_t}

Consider the first difference:

$$\Delta Y_{t+1} = Y_{t+1} - Y_t = (U_{t+1} - U_t) + (V_{t+1} - V_t) \;=\; \eta_{t+1} + V_{t+1} - V_t. \tag{5.24}$$

This equation implies that we may replace V_{t+1} by V_t, η_{t+1} and ΔY_{t+1} as

$$V_{t+1} = V_t - \eta_{t+1} + \Delta Y_{t+1}. \tag{5.25}$$

Consider the following first difference for the next time period:

$$\begin{aligned} \Delta Y_{t+2} &= Y_{t+2} - Y_{t+1} \\ &= \eta_{t+2} + V_{t+2} - V_{t+1} \;=\; (\rho_{t+2} - 1) V_{t+1} + G_{t+2}(\varepsilon_{t+2}, \varepsilon_{t+1}) + \eta_{t+2}. \end{aligned} \tag{5.26}$$

Replacing V_{t+1} by the expression in equation (5.25), we obtain

$$\frac{\Delta Y_{t+2}}{\rho_{t+2} - 1} - \Delta Y_{t+1} \;=\; V_t + \frac{G_{t+2}(\varepsilon_{t+2}, \varepsilon_{t+1}) + \eta_{t+2}}{\rho_{t+2} - 1} - \eta_{t+1} \;\equiv\; V_t + e_{t+1}. \tag{5.27}$$

With the pair of equations (5.20) and (5.27), we obtain two measurements, $\frac{\Delta Y_{t+2}}{\rho_{t+2}-1} - \Delta Y_{t+1}$ and Y_t up to an unknown scalar parameter ρ_{t+2}, of the latent variable V_t with classical measurement errors, U_t and e_{t+1}, satisfying the mutual independence among V_t, U_t and e_{t+1}. By Kotlarski's identity , the distribution of V_t is identified up to the unknown scalar parameter ρ_{t+2} as

$$\begin{aligned} f_{V_t}(v) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\tau v} \phi_{V_t}(\tau) d\tau, \qquad \text{where } i = \sqrt{-1} \\ \phi_{V_t}(\tau) &= \exp\left[\int_0^{\tau} \frac{iE\left[\left(\frac{\Delta Y_{t+2}}{\rho_{t+2}-1} - \Delta Y_{t+1} \right) \exp(isY_t) \right]}{E[\exp(isY_t)]} ds \right]. \end{aligned} \tag{5.28}$$

For the current step, a well-definition of the last identifying formula requires the following non-unit root assumption for the transitory state.

Assumption 5.3.2 *$\rho_t \neq 1$ for all t.*

Step 2: Identification of ρ_t

The previous step shows identification of f_{V_t} up to the unknown scalar parameter ρ_t. We now discuss alternative routes of identifying the AR parameter ρ_t. Combining (5.23) and (5.26), we obtain

$$\begin{aligned} \Delta Y_{t+2} &= (\rho_{t+2}-1) V_{t+1} + G_{t+2}(\varepsilon_{t+2}, \varepsilon_{t+1}) + \eta_{t+2} \\ &= (\rho_{t+2}-1)(\rho_{t+1} V_t + G_{t+1}(\varepsilon_{t+1}, \varepsilon_t)) + G_{t+2}(\varepsilon_{t+2}, \varepsilon_{t+1}) + \eta_{t+2} \end{aligned} \tag{5.29}$$

Eliminating V_t with (5.27) yields

$$\begin{aligned} & \frac{\Delta Y_{t+2}}{(\rho_{t+2}-1)\rho_{t+1}} - \left(\frac{\Delta Y_{t+2}}{\rho_{t+2}-1} - \Delta Y_{t+1} \right) \\ = & \frac{(\rho_{t+2}-1) G_{t+1}(\varepsilon_{t+1}, \varepsilon_t) + G_{t+2}(\varepsilon_{t+2}, \varepsilon_{t+1}) + \eta_{t+2}}{(\rho_{t+2}-1)\rho_{t+1}} - \left(\frac{G_{t+2}(\varepsilon_{t+2}, \varepsilon_{t+1}) + \eta_{t+2}}{\rho_{t+2}-1} - \eta_{t+1} \right). \end{aligned}$$

Notice that the last expression is independent of $Y_{t-1} = V_{t-1} + U_{t-1}$ under Assumption 6.1.4, and we get the moment restriction

$$cov\left(\left(\frac{1-\rho_{t+1}}{\rho_{t+1}(1-\rho_{t+2})} \Delta Y_{t+2} - \Delta Y_{t+1} \right), Y_{t-1} \right) = 0. \tag{5.30}$$

For a better view, we rewrite it as

$$\rho_{t+1} \frac{1-\rho_{t+2}}{1-\rho_{t+1}} = \frac{cov(\Delta Y_{t+2}, Y_{t-1})}{cov(\Delta Y_{t+1}, Y_{t-1})}. \tag{5.31}$$

We can see from this equation that, by imposing one restriction on the sequence $\rho_{t+1}, \rho_{t+2}, \cdots$, we can sequentially identify these AR parameters. Examples of such a restriction include

$$\begin{aligned} \rho_{t+1} &= \text{a known constant}, \quad \text{or} \\ \rho_{t+1} &= \rho_{t+2}. \end{aligned}$$

In the former case, one can recursively identify $\rho_{t+2}, \rho_{t+3}, \cdots$ by iterating (5.31). In the latter case, (5.31) directly yields the identifying formula

$$\rho_{t+1} = \frac{cov(\Delta Y_{t+2}, Y_{t-1})}{cov(\Delta Y_{t+1}, Y_{t-1})}, \tag{5.32}$$

provided that $cov(\Delta Y_{t+1}, Y_{t-1}) \neq 0$ and Assumption 5.3.2. We state this restriction as an assumption below.

Assumption 5.3.3 *$cov(\Delta Y_{t+1}, Y_{t-1}) \neq 0$ and $\rho_{t+1} = \rho_{t+2}$ for all t.*

Step 3: Identification of f_{η_t}, $f_{U_2,\cdots,U_{T-2}}$ and $f_{V_2,\cdots,V_{T-2}}$

Steps 1 and 2 identify the characteristic function ϕ_{V_t} by (5.28) for $t = 2, \cdots, T-2$. Given that U_t and V_t are independent, we identify the marginal distribution of U_t via the deconvolution:

$$\phi_{U_t} = \frac{\phi_{Y_t}}{\phi_{V_t}}. \tag{5.33}$$

Similarly and consequently, we also identify the marginal distribution of η_t by

$$\phi_{\eta_t} = \frac{\phi_{U_t}}{\phi_{U_{t-1}}}. \tag{5.34}$$

Notice that the independence between the permanent state U_{t-1} and the innovation η_t implies that

$$f_{U_t|U_{t-1}}(u_t, u_{t-1}) = f_{\eta_t}(u_t - u_{t-1}) \tag{5.35}$$

holds. Therefore, the joint distribution of $(U_2, U_3, \cdots, U_{T-2})$ is identified by

$$f_{U_2,U_3,\cdots,U_{T-2}} = f_{U_{T-2}|U_{T-3}} f_{U_{T-3}|U_{T-4}} \cdots f_{U_3|U_2} f_{U_2}. \tag{5.36}$$

Moreover, the independence between the process $\{U_t\}$ and the process $\{V_t\}$ implies

$$\phi_{Y_2,\cdots,Y_{T-2}} = \phi_{U_2,\cdots,U_{T-2}} \phi_{V_2,\cdots,V_{T-2}},$$

where $\phi_{Y_2,\cdots,Y_{T-2}}$ is the joint characteristic function of $Y_2, \cdots, Y_{T-2}$. Therefore, the joint distribution of the transitory states $(V_2, \cdots, V_{T-2})$ is also identified from the corresponding joint characteristic function

$$\phi_{V_2,\cdots,V_{T-2}} = \frac{\phi_{Y_2,\cdots,Y_{T-2}}}{\phi_{U_2,\cdots,U_{T-2}}}. \tag{5.37}$$

This step requires the following assumption.

Assumption 5.3.4 *(i)* $\phi_{U_1,\cdots,U_T}(s_1, \cdots, s_T) = E[\exp(is_1U_1 + \cdots + is_TU_T)]$ *is not equal to zero for any real* $(s_1, \cdots, s_T)$*. (ii) For each of* $(Y_1, \cdots, Y_T)$*,* $(U_1, \cdots, U_T)$*,* $(V_1, \cdots, V_T)$*,* η_t *and* ε_t*, the marginal and joint distributions are absolutely continuous with respect to the Lebesgue measure, and the marginal and joint characteristic functions are absolutely integrable.*

Part (i) of this assumption is the assumption of non-vanishing characteristic function as in Li and Vuong (1998) with a multivariate extension. It corresponds to the "completeness" assumption for nonparametric identification as in Hu and Schennach (2008) and Arellano, Blundell, and Bonhomme (2017) – see also D'Haultfoeuille (2011). In the univariate context, this assumption is known to be satisfied by most of the popular continuous distribution families, while counter-examples of distribution families violating this assumption are the uniform, the truncated normal, and many discrete distributions (Evdokimov and White, 2012). Similar remarks apply to multivariate distribution families, though there are not

many stylized families of multivariate distributions. Particularly, the assumption is satisfied by the multivariate normal distributions.

We summarize the results as follows.

Proposition 1 *Suppose that Assumptions 6.1.4, 5.3.2, 6.1.6, and 5.3.4 hold. The joint distribution of* $(Y_1, \cdots, Y_T)$ *uniquely determines the marginal distribution of* η_t *for* $t = 3, 4, \cdots, T-2$*, the joint distribution of* $(U_2, \cdots, U_{T-2})$*, and the joint distribution of* $(V_2, \cdots, V_{T-2})$*, together with* ρ_t *for* $t = 3, 4, \cdots, T$*.*

Step 4: Identification of f_{ε_t} and G_t

Since G_t is arbitrarily nonparametric, we cannot identify the nonparametric distribution of ε_t in general. However, we may identify its distribution if the following restriction is imposed.

Assumption 5.3.5 *The MA function* G_t *takes the form* $G_t(\varepsilon_t, \varepsilon_{t-1}) = \varepsilon_t + g_t(\varepsilon_{t-1})$ *with the location normalizations* $E[\varepsilon_t] = E[g_t(\varepsilon_{t-1})] = 0$*.*

Since we have identified ρ_t for $t = 3, 4, \cdots, T$ and the joint distribution $f_{V_2, \cdots, V_{T-2}}$, we identify the joint distribution of two composite random variables $(V_t - \rho_t V_{t-1})$ and $(V_{t-1} - \rho_{t-1} V_{t-2})$. These two random variables can be in turn rewritten as follows:

$$\begin{aligned} V_t - \rho_t V_{t-1} &= \varepsilon_t + g_t(\varepsilon_{t-1}) \\ V_{t-1} - \rho_{t-1} V_{t-2} &= \varepsilon_{t-1} + g_{t-1}(\varepsilon_{t-2}) \end{aligned} \tag{5.38}$$

The three shocks to the transitory states on the right-hand side are mutually independent. When the function $g_t(x) = \lambda_t x$ is linear, Reiersol (1950) shows that the coefficient λ_t is generally identified if ε_t is not normally distributed. Schennach and Hu (2013) generalize this result to nonlinear cases. We may identify the function g_t for $t = 4, \cdots, T-2$ and the marginal distribution of ε_t for $t = 3, \cdots, T-2$ using the results in Schennach and Hu (2013).

Assumption 5.3.6 (Schennach and Hu (2013)) *(i) The marginal characteristic functions of* ε_{t-1}*,* ε_t*,* $g_t(\varepsilon_{t-1})$*, and* $g_{t-1}(\varepsilon_{t-2})$ *do not vanish on the real line. (ii) The density function* $f_{\varepsilon_{t-1}}$ *of* ε_{t-1} *exists and is uniformly bounded. (iii)* g_t *is continuously differentiable, strictly monotone, and is not exactly of the form* $g_t(\varepsilon_{t-1}) = a + b\ln(e^{c\varepsilon_{t-1}} + d)$ *for* $a, b, c, d \in \mathbb{R}$*.*

This assumption states Assumptions 1–6 and an additional condition of Theorem 1 in Schennach and Hu (2013) in terms of our notation. (The notations in Schennach and Hu (2013) and our notations are reconciled by $y := V_t - \rho_t V_{t-1}$, $x := V_{t-1} - \rho_{t-1} V_{t-2}$, $x^* := \varepsilon_{t-1}$, $\Delta y := \varepsilon_t$, $\Delta x := g_{t-1}(\varepsilon_{t-2})$ and $g := g_t$.) The first part of Assumption 1 in Schennach and Hu (2013) is implied by our Assumption 6.1.4 (ii), and hence is not included

in our Assumption 5.3.6. Likewise, the second part of Assumption 1 in Schennach and Hu (2013) is implied by our Assumption 6.1.7, and hence is not included in Assumption 5.3.6. Part (i) is similar to Assumption 5.3.4 (i). As discussed earlier, it corresponds to the "completeness" assumption for nonparametric identification (D'Haultfoeuille, 2011). This assumption is known to be satisfied by most of the popular continuous distribution families, while counter-examples of distribution families violating this assumption are the uniform, the truncated normal, and many discrete distributions (Evdokimov and White, 2012). Part (ii) of the assumption is also satisfied by most of the popular continuous distribution families, with the chi-square distribution of one degree of freedom being a major counter-example. Part (iii) is a set of requirement for the function g_t in the MA decomposition.

Proposition 2 *Suppose that Assumption 6.1.7 and 5.3.6, in addition to the assumptions in Proposition 1, are satisfied. The joint distribution of* $(Y_1, \cdots, Y_T)$ *uniquely determines the marginal distribution of* ε_t *and the MA function* G_t*.*

This result guarantees nonparametric identification but the identification is not constructive and therefore a plug-in estimator is not available. A closed-form estimator is available at the cost of further assuming the linear MA structure as in Reiersol (1950):

$$g_t(x) = \lambda_t x.$$

In this case, (5.38) simplifies to the classical repeated measurement model:

$$\begin{aligned} V_t - \rho_t V_{t-1} &= \varepsilon_t + \lambda_t \varepsilon_{t-1} \\ V_{t-1} - \rho_{t-1} V_{t-2} &= \varepsilon_{t-1} + \lambda_{t-1} \varepsilon_{t-2} \end{aligned}$$

Therefore, we may use Kotlarski's identity to obtain the closed-form identifying formula

$$\begin{aligned} f_{\varepsilon_t}(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\tau x} \phi_{\varepsilon_t}(\tau) d\tau, \qquad \text{where} \\ \phi_{\varepsilon_t}(\tau) &= \exp\left[\int_0^{\tau} \frac{iE\left[\left(\frac{V_{t+1} - \rho_{t+1} V_t}{\lambda_{t+1}} \right) \exp\left(is\left(V_t - \rho_t V_{t-1} \right) \right) \right]}{E\left[\exp\left(V_t - \rho_t V_{t-1} \right) \right]} ds \right] \end{aligned} \tag{5.39}$$

where the expectations can be computed using the closed-form identifying formula (5.37) for the joint distribution of (V_{t-1}, V_t, V_{t+1}) obtained in the previous step.

To compute the closed form (5.39) it remains to identify the unknown scalar λ_{t+1}. We can find the moment restrictions

$$\begin{aligned} var(\varepsilon_{t+1}) + \lambda_{t+1}^2 var(\varepsilon_t) &= var(V_{t+1} - \rho_{t+1} V_t) \\ \lambda_{t+1} var(\varepsilon_t) &= cov(V_t - \rho_t V_{t-1}, V_{t+1} - \rho_{t+1} V_t) \\ \lambda_{t+2} var(\varepsilon_{t+1}) &= cov(V_{t+1} - \rho_{t+1} V_t, V_{t+2} - \rho_{t+2} V_{t+1}) \end{aligned} \tag{5.40}$$

where the values on the right-hand sides can be computed again using the closed-form identifying formula (5.37) for the joint distribution of $(V_{t-1}, V_t, V_{t+1}, V_{t+2})$ obtained in the

previous step. The left-hand sides contain four unknowns, $var(\varepsilon_t)$, $var(\varepsilon_{t+1})$, λ_{t+1} and λ_{t+2}. Therefore, one restriction is necessary for identification of λ_{t+1} using the above three equations.

6

Applications in Structural Econometrics

6.1 Dynamic Discrete Choice with Unobserved State Variables

Hu and Shum (2012) show that the transition kernel of a Markov process $\{W_t, X_t^*\}$ can be uniquely determined by the joint distribution of four periods of data $\{W_{t+1}, W_t, W_{t-1}, W_{t-2}\}$. This result can be directly applied to identification of dynamic discrete choice model with unobserved state variables. Such a Markov process may characterize the optimal path of the decision and the state variables in Markov dynamic optimization problems. Let $W_t = (Y_t, M_t)$, where Y_t is the agent's choice in period t, and M_t denotes the period-t observed state variable, while X_t^* is the unobserved state variable. For Markovian dynamic optimization models, the transition kernel can be decomposed as follows:

$$f_{W_t, X_t^* | W_{t-1}, X_{t-1}^*} = f_{Y_t | M_t, X_t^*} f_{M_t, X_t^* | Y_{t-1}, M_{t-1}, X_{t-1}^*}. \tag{6.1}$$

The first term on the right hand side is the conditional choice probability for the agent's optimal choice in period t. The second term is the joint law of motion of the observed and unobserved state variables. As shown in Hotz and Miller (1993) , the identified Markov law of motion may be a crucial input in the estimation of Markovian dynamic models. One advantage of this conditional choice probability approach is that a parametric specification of the model leads to a parametric GMM estimator. That implies an estimator for a dynamic discrete choice model with unobserved state variables, where one can identify the Markov transition kernel containing unobserved state variables, and then apply the conditional choice probability estimator to estimate the model primitives. Hu and Shum (2013) extend this result to dynamic games with unobserved state variables.

6.1.1 Background

In this paper, we consider the identification of a Markov process $\{W_t, X_t^*\}$ when only $\{W_t\}$, a subset of the variables, is observed. In structural dynamic models, W_t typically consists of the choice variables and observed state variables of an optimizing agent. X_t^* denotes time-varying serially correlated unobserved state variables (or agent-specific unobserved heterogeneity), which are observed by the agent, but not by the econometrician.

We demonstrate two main results. First, in the non-stationary case, where the Markov law of motion $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$, can vary across periods t, we show that, for any period t, $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$ is identified from five periods of data $W_{t+1},\ldots,W_{t-3}$. Second, in the stationary case, where $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$ is the same across all t, only four observations $W_{t+1},\ldots,W_{t-2}$, for some t, are required for identification.

In most applications, W_t consists of two components $W_t = (Y_t, M_t)$, where Y_t denotes the agent's action in period t, and M_t denotes the period-t observed state variable(s). X_t^* are time-varying unobserved state variables (USV for short), which are observed by agents and affect their choice of Y_t, but unobserved by the econometrician. The economic importance of models with unobserved state variables has been recognized since the earliest papers on the structural estimation of dynamic optimization models. Two examples are:

[1] **Miller's 1984** job matching model was one of the first empirical dynamic discrete choice models with unobserved state variables. Y_t is an indicator for the occupation chosen by a worker in period t, and the unobserved state variables X_t^* are the time-varying posterior means of workers' beliefs regarding their occupation-specific match values. The observed state variables M_t include job tenure and education level. [2] **Pakes (1986)** estimates an optimal stopping model of the year-by-year renewal decision on European patents. In his model, the decision variable Y_t is an indicator for whether a patent is renewed in year t, and the unobserved state variable X_t^* is the profitability from the patent in year t, which varies across years and is not observed by the econometrician. The observed state variable M_t could be other time-varying factors, such as the stock price or total sales of the patent-holding firm, which affect the renewal decision. ■

These two early papers demonstrated that dynamic optimization problems with an unobserved process partly determining the state variables are indeed empirically tractable. Their authors (cf. (Miller, 1984, section V); Pakes and Simpson (1989)) also provided some discussion of the restrictions implied on the data by their models, thus highlighting how identification has been a concern since the earliest structural empirical applications of dynamic models with unobserved state variables. Obviously, the nonparametric identification of these complex nonlinear models has important practical relevance for empirical researchers, and our goal here is to provide identification results which apply to a broad class of Markovian dynamic models with time-varying unobserved state variables.

Our main result concerns the identification of the Markov law of motion $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$. Once this is known, it factors into conditional and marginal distributions of economic interest. For Markovian dynamic optimization models (such as the examples given above),

the law of motion $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$ factors into

$$\begin{aligned} f_{W_t,X_t^*|W_{t-1},X_{t-1}^*} &= f_{Y_t,M_t,X_t^*|Y_{t-1},M_{t-1},X_{t-1}^*} \\ &= \underbrace{f_{Y_t|M_t,X_t^*}}_{\text{CCP}} \cdot \underbrace{f_{M_t,X_t^*|Y_{t-1},M_{t-1},X_{t-1}^*}}_{\text{state law of motion}}. \end{aligned} \tag{6.2}$$

The first term denotes the conditional choice probability for the agent's optimal choice in period t. The second term is the Markovian law of motion for the state variables (M_t, X_t^*).

Once the CCP's and the law of motion for the state variables are recovered, it is straightforward to use them as inputs in a CCP-based approach for estimating dynamic discrete-choice models. This approach was pioneered in Hotz and Miller (1993) and Hotz et al. (1994).[1] A general criticism of these methods is that they cannot accommodate unobserved state variables. In response, Aguirregabiria and Mira (2007), Buchinsky et al. (2004), and Houde and Imai (2006), among others, recently developed CCP-based estimation methodologies allowing for agent-specific unobserved heterogeneity, which is the special case where the latent X_t^* is time-invariant. Arcidiacono and Miller (2011) developed a CCP-based approach to estimate dynamic discrete models where X_t^* varies over time according to an exogenous first-order discrete Markov process.[2]

While these papers have focused on estimation, our focus is on identification. Our identification approach is novel because it is based on recent econometric results in nonlinear measurement error models. Specifically, we show that the identification results in Hu and Schennach (2008) and Carroll et al. (2010) for nonclassical measurement models (where the measurement error is not assumed to be independent of the latent "true" variable) can be applied to Markovian dynamic models, and we use those results to establish nonparametric identification.

Our results extend nonparametric identification to classes of models not covered in the existing identification literature. When the unobserved state variable X_t^* is discrete, our results cover cases where X_t^* is time-varying and can evolve depending on past values of the observed variables W_{t-1}. This is new in the literature. When X_t^* is continuous, however, our identification results require high-level "completeness" assumptions which are difficult to verify in practice. One worked-out example (in Section 4.2) shows that these completeness assumptions are implied by independent initial conditions, in addition to other restrictions on the laws of motion of the state variables: while this is new ground, these restrictions are nevertheless strong. Because of this, when X_t^* is continuous, we see our results more as a useful starting point, rather than a final word on the subject.

Kasahara and Shimotsu (2009) (hereafter KS) consider the identification of dynamic models with discrete unobserved heterogeneity, where the latent variable $X_t^* = X^*$ is time-

[1] Subsequent methodological developments for CCP-based estimation include Aguirregabiria and Mira Aguirregabiria and Mira (2002), Aguirregabiria and Mira (2007), Pesendorfer and Schmidt-Dengler (2008), Bajari et al. (2007a), Pakes et al. (2007), and Hong and Shum (2007).At the same time, Magnac and Thesmar (2002) and Bajari et al. (2007b) use the CCP logic to provide identification results for dynamic discrete-choice models.

[2] That is, X_t^* is discrete-valued, and depends stochastically only on X_{t-1}^*, and not on any other variables. We relax this in Section 4.1 below.

invariant and discrete. KS demonstrate that the Markov law of motion $W_{t+1}|W_t, X^*$ is identified in this setting, using six periods of data. Relative to this, we consider a more general setting where the unobserved X_t^* is allowed to vary over time (as in the Miller and Pakes examples above), and can evolve depending on past values of the observed variables W_{t-1}.

Henry et al. (2008) (hereafter HKS) exploit exclusion restrictions to identify Markov regime-switching models with a discrete and latent state variable. While our identification arguments are quite distinct from those in HKS, our results share some of HKS's intuition, because we also exploit the feature of first-order Markovian models that, conditional on W_{t-1}, W_{t-2} is an "excluded variable" which affects W_t only via the unobserved state X_t^*.[3]

Cunha et al. (2010) apply the result of Hu and Schennach (2008) to show nonparametric identification of a nonlinear factor model consisting of $(W_t, W_t', W_t'', X_t^*)$, where the observed processes $\{W_t\}_{t=1}^T$, $\{W_t'\}_{t=1}^T$, and $\{W_t''\}_{t=1}^T$ constitute noisy measurements of the latent process $\{X_t^*\}_{t=1}^T$, contaminated with random disturbances. In contrast, we consider a setting where (W_t, X_t^*) jointly evolves as a dynamic Markov process. We use observations of W_t in different periods t to identify the conditional density of $(W_t, X_t^*|W_{t-1}, X_{t-1}^*)$. Thus, our model and identification strategy differ from theirs.

6.1.2 The Discrete Case

Identification in the Discrete Case

We start with the case where the unobserved state variable is discrete. Let (W_t, X_t^*) denote a bivariate discrete first-order Markov process where W_t and X_t^* are both scalars sharing the same support $\mathcal{W}_t \equiv \{1, 2, ..., K\}$. The identification results below can be straightforwardly extended to the case where W_t has more possible values than X_t^*. We assume

Assumption 6.1.1 *Suppose W_t and X_t^* share the same support $\mathcal{W}_t \equiv \{1, 2, ..., K\}$ and the dynamic process $\{W_t, X_t^*\}$ satisfy (i)* First-order Markov: $f_{W_t, X_t^*|W_{t-1}, X_{t-1}^*, \Omega_{<t-1}} = f_{W_t, X_t^*|W_{t-1}, X_{t-1}^*}$, *where $\Omega_{<t-1} \equiv \{W_{t-2}, ..., W_1, X_{t-2}^*, ..., X_1^*\}$, the history up to (but not including) $t-1$.*
(ii) Limited feedback: $f_{W_t|W_{t-1}, X_t^*, X_{t-1}^*} = f_{W_t|W_{t-1}, X_t^*}$.

[3] Similarly, Bouissou et al. (1986) exploit the Markov restrictions on a stochastic process X to formulate tests for the noncausality of another process Y on X.

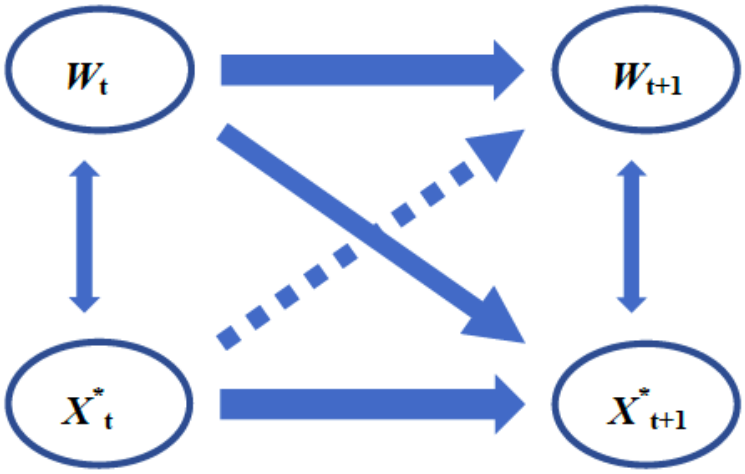


Figure 6.1: Limited-feedback assumption

Assumption 6.1.1 implies that,

$$
\begin{aligned}
& f_{W_{t+1},W_t,W_{t-1},W_{t-2}} \\
= & \sum_{X_t^*} \sum_{X_{t-1}^*} f_{W_{t+1},W_t,X_t^*,X_{t-1}^*,W_{t-1},W_{t-2}} \\
= & \sum_{X_t^*} \sum_{X_{t-1}^*} f_{W_{t+1}|W_t,W_{t-1},W_{t-2},X_t^*,X_{t-1}^*} f_{W_t,X_t^*|W_{t-1},W_{t-2},X_{t-1}^*} f_{X_{t-1}^*,W_{t-1},W_{t-2}} \\
= & \sum_{X_t^*} \sum_{X_{t-1}^*} f_{W_{t+1}|W_t,X_t^*} f_{W_t,X_t^*|W_{t-1},X_{t-1}^*} f_{X_{t-1}^*,W_{t-1},W_{t-2}} \\
= & \sum_{X_t^*} \sum_{X_{t-1}^*} f_{W_{t+1}|W_t,X_t^*} f_{W_t|W_{t-1},X_t^*,X_{t-1}^*} f_{X_t^*|W_{t-1},X_{t-1}^*} f_{X_{t-1}^*,W_{t-1},W_{t-2}} \\
= & \sum_{X_t^*} \sum_{X_{t-1}^*} f_{W_{t+1}|W_t,X_t^*} f_{W_t|W_{t-1},X_t^*,X_{t-1}^*} f_{X_t^*|W_{t-1},W_{t-2},X_{t-1}^*} f_{X_{t-1}^*,W_{t-1},W_{t-2}} \\
= & \sum_{X_t^*} \sum_{X_{t-1}^*} f_{W_{t+1}|W_t,X_t^*} f_{W_t|W_{t-1},X_t^*,X_{t-1}^*} f_{X_t^*,X_{t-1}^*,W_{t-1},W_{t-2}}.
\end{aligned}
$$

Assumption 6.1.4(ii) then implies

$$
\begin{aligned}
f_{W_{t+1},W_t,W_{t-1},W_{t-2}} &= \sum_{X_t^*} f_{W_{t+1}|W_t,X_t^*} f_{W_t|W_{t-1},X_t^*} \left(\sum_{X_{t-1}^*} f_{X_t^*,X_{t-1}^*,W_{t-1},W_{t-2}} \right) \\
&= \sum_{X_t^*} f_{W_{t+1}|W_t,X_t^*} f_{W_t|W_{t-1},X_t^*} f_{X_t^*,W_{t-1},W_{t-2}}. \qquad (6.3)
\end{aligned}
$$

For any $(w_t, w_{t-1}) \in \mathcal{W}_t \times \mathcal{W}_{t-1}$, we define matrices as follows,

$$\begin{aligned}
M_{W_{t+1},w_t,w_{t-1},W_{t-2}} &= \left[f_{W_{t+1},W_t,W_{t-1},W_{t-2}}\left(i,w_t,w_{t-1},j\right)\right]_{i=1,2,...,K;j=1,2,...,K} \\
M_{W_{t+1}|w_t,X_t^*} &= \left[f_{W_{t+1}|W_t,X_t^*}\left(i|w_t,j\right)\right]_{i=1,2,...,K;j=1,2,...,K} \\
D_{w_t|w_{t-1},X_t^*} &= \begin{bmatrix} f_{W_t|W_{t-1},X_t^*}\left(w_t|w_{t-1},1\right) & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & f_{W_t|W_{t-1},X_t^*}\left(w_t|w_{t-1},K\right) \end{bmatrix} \\
M_{X_t^*,w_{t-1},W_{t-2}} &= \left[f_{X_t^*,W_{t-1},W_{t-2}}\left(i,w_{t-1},j\right)\right]_{i=1,2,...,K;j=1,2,...,K}
\end{aligned}$$

In general, we define a matrix representation of a probability distribution as follows: for discrete random variables R_1, R_2, R_3, the $(i+1, j+1)$-th element of the matrix M_{R_1,r_2,R_3} contains the joint probability that $(R_1 = i, R_2 = r_2, R_3 = j)$, for $i, j \in \{1, 2, ..., K\}$. Equation (6.3) is then equivalent to

$$M_{W_{t+1},w_t,w_{t-1},W_{t-2}} = M_{W_{t+1}|w_t,X_t^*} D_{w_t|w_{t-1},X_t^*} M_{X_t^*,w_{t-1},W_{t-2}} \ . \tag{6.4}$$

Notice that for fixed (w_t, w_{t-1}), we only have two measurements of the latent X_t^*. However, an important observation is that for (w_t, w_{t-1}),

$$M_{W_{t+1},w_t,w_{t-1},W_{t-2}} = \underbrace{M_{W_{t+1}|w_t,X_t^*}}_{\text{no } w_{t-1}} \ \underbrace{D_{w_t|w_{t-1},X_t^*}}_{\text{only } K \text{ unkowns.}} \ \underbrace{M_{X_t^*,w_{t-1},W_{t-2}}}_{\text{no } w_t}$$

Therefore, we may consider different values of (W_t, W_{t-1}) as follows: for (w_t, w_{t-1}), $(\overline{w}_t, w_{t-1})$, $(\overline{w}_t, \overline{w}_{t-1})$ $(w_t, \overline{w}_{t-1})$,

$$\begin{aligned}
M_{W_{t+1},w_t,w_{t-1},W_{t-2}} &= M_{W_{t+1}|w_t,X_t^*} \quad D_{w_t|w_{t-1},X_t^*} \quad \underbrace{M_{X_t^*,w_{t-1},W_{t-2}}}_{\| \ \overline{M_{X_t^*,w_{t-1},W_{t-2}}}} \\
M_{W_{t+1},\overline{w}_t,w_{t-1},W_{t-2}} &= \underbrace{M_{W_{t+1}|\overline{w}_t,X_t^*}}_{\| \ \overline{M_{W_{t+1}|\overline{w}_t,X_t^*}}} \quad D_{\overline{w}_t|w_{t-1},X_t^*} \quad \overline{M_{X_t^*,w_{t-1},W_{t-2}}} \\
M_{W_{t+1},\overline{w}_t,\overline{w}_{t-1},W_{t-2}} &= \overline{M_{W_{t+1}|\overline{w}_t,X_t^*}} \quad D_{\overline{w}_t|\overline{w}_{t-1},X_t^*} \quad \underbrace{M_{X_t^*,\overline{w}_{t-1},W_{t-2}}}_{\| \ \overline{M_{X_t^*,\overline{w}_{t-1},W_{t-2}}}} \\
M_{W_{t+1},w_t,\overline{w}_{t-1},W_{t-2}} &= M_{W_{t+1}|w_t,X_t^*} \quad D_{w_t|\overline{w}_{t-1},X_t^*} \quad \overline{M_{X_t^*,\overline{w}_{t-1},W_{t-2}}}
\end{aligned}$$

Under the assumption that the four matrices on the LHS are invertible, which is directly testable, we may have

$$\begin{aligned}
\mathbf{A} &\equiv M_{W_{t+1},w_t,w_{t-1},W_{t-2}} M^{-1}_{W_{t+1},\overline{w}_t,w_{t-1},W_{t-2}} \\
&= M_{W_{t+1}|w_t,X_t^*} D_{w_t|w_{t-1},X_t^*} D^{-1}_{\overline{w}_t|w_{t-1},X_t^*} M^{-1}_{W_{t+1}|\overline{w}_t,X_t^*}.
\end{aligned}$$

Similar manipulations lead to

$$\begin{aligned}\mathbf{B} &\equiv M_{W_{t+1},\overline{w}_t,\overline{w}_{t-1},W_{t-2}} M^{-1}_{W_{t+1},w_t,\overline{w}_{t-1},W_{t-2}} \\ &= M_{W_{t+1}|\overline{w}_t,X_t^*} D_{\overline{w}_t|\overline{w}_{t-1},X_t^*} D^{-1}_{w_t|\overline{w}_{t-1},X_t^*} M^{-1}_{W_{t+1}|w_t,X_t^*}.\end{aligned}$$

Assumption 6.1.5(i) guarantees that, for any w_t, ($\overline{w}_t$, w_{t-1}, $\overline{w}_{t-1}$) exist so that matrices $\mathbf{A}$ and $\mathbf{B}$ exist. Finally, we obtain

$$\begin{aligned}\mathbf{AB} &= M_{W_{t+1}|w_t,X_t^*} D_{w_t|w_{t-1},X_t^*} D^{-1}_{\overline{w}_t|w_{t-1},X_t^*} \left(M^{-1}_{W_{t+1}|\overline{w}_t,X_t^*} M_{W_{t+1}|\overline{w}_t,X_t^*} \right) \times \\ &\quad \times D_{\overline{w}_t|\overline{w}_{t-1},X_t^*} D^{-1}_{w_t|\overline{w}_{t-1},X_t^*} M^{-1}_{W_{t+1}|w_t,X_t^*} \\ &= M_{W_{t+1}|w_t,X_t^*} \left(D_{w_t|w_{t-1},X_t^*} D^{-1}_{\overline{w}_t|w_{t-1},X_t^*} D_{\overline{w}_t|\overline{w}_{t-1},X_t^*} D^{-1}_{w_t|\overline{w}_{t-1},X_t^*} \right) M^{-1}_{W_{t+1}|w_t,X_t^*} \\ &\equiv M_{W_{t+1}|w_t,X_t^*} D_{w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},X_t^*} M^{-1}_{W_{t+1}|w_t,X_t^*}, \qquad (6.5)\end{aligned}$$

where

$$\begin{aligned}D_{w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},X_t^*} &= D_{w_t|w_{t-1},X_t^*} D^{-1}_{\overline{w}_t|w_{t-1},X_t^*} D_{\overline{w}_t|\overline{w}_{t-1},X_t^*} D^{-1}_{w_t|\overline{w}_{t-1},X_t^*} \\ &= \begin{bmatrix} k(w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},1) & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & k(w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},K) \end{bmatrix} \qquad (6.6)\end{aligned}$$

with

$$k(w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},x_t^*) \equiv \frac{f_{W_t|W_{t-1},X_t^*}(w_t|w_{t-1},x_t^*) f_{W_t|W_{t-1},X_t^*}(\overline{w}_t|\overline{w}_{t-1},x_t^*)}{f_{W_t|W_{t-1},X_t^*}(\overline{w}_t|w_{t-1},x_t^*) f_{W_t|W_{t-1},X_t^*}(w_t|\overline{w}_{t-1},x_t^*)}$$

This equation implies that the observed matrix $\mathbf{AB}$ on the left hand side has an inherent eigenvalue-eigenfunction decomposition, with the eigenvalues corresponding to the function $k(w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},x_t^*)$ and the eigenfunctions corresponding to the probability function $f_{W_{t+1}|W_t,X_t^*}(\cdot|w_t,x_t^*)$. Such a decomposition is similar to the decomposition in Hu (2008). Under a similar set of assumptions as in Hu (2008), such as distinctive eigenvalues and ordering assumptions (6.1.6 and 6.1.7), we may achieve a unique decomposition and identify $f_{W_{t+1}|W_t,X_t^*}$ or equivalently $M_{W_{t+1}|w_t,X_t^*}$ for all w_t.

Next, we show that identification of $f_{W_{t+1}|W_t,X_t^*}$ leads to identification of the Markov kernel $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$. The Markov properties implies

$$\begin{aligned}f_{W_{t+1},W_t,W_{t-1},W_{t-2}} &= \sum_{X_t^*} f_{W_{t+1}|W_t,X_t^*} f_{W_t,X_t^*,W_{t-1},W_{t-2}} \\ f_{W_t,X_t^*,W_{t-1},W_{t-2}} &= \sum_{X_{t-1}^*} f_{W_t,X_t^*|W_{t-1},X_{t-1}^*} f_{X_{t-1}^*,W_{t-1},W_{t-2}}. \qquad (6.7)\end{aligned}$$

In matrix notation, for fixed w_t, w_{t-1}, the above equations are expressed:

$$\begin{aligned}M_{W_{t+1},w_t,w_{t-1},W_{t-2}} &= M_{W_{t+1}|w_t,X_t^*} M_{w_t,X_t^*,w_{t-1},W_{t-2}} \\ M_{w_t,X_t^*,w_{t-1},W_{t-2}} &= M_{w_t,X_t^*|w_{t-1},X_{t-1}^*} M_{X_{t-1}^*,w_{t-1},W_{t-2}}.\end{aligned}$$

Substituting the second line into the first, we get

$$\begin{aligned} M_{W_{t+1},w_t,w_{t-1},W_{t-2}} &= M_{W_{t+1}|w_t,X_t^*} M_{w_t,X_t^*|w_{t-1},X_{t-1}^*} M_{X_{t-1}^*,w_{t-1},W_{t-2}} \\ \Leftrightarrow M_{w_t,X_t^*|w_{t-1},X_{t-1}^*} M_{X_{t-1}^*,w_{t-1},W_{t-2}} &= M_{W_{t+1}|w_t,X_t^*}^{-1} M_{W_{t+1},w_t,w_{t-1},W_{t-2}}. \end{aligned} \tag{6.8}$$

We then eliminate $M_{X_{t-1}^*,w_{t-1},W_{t-2}}$ from the above. We have

$$f_{W_t,W_{t-1},W_{t-2}} = \sum_{X_{t-1}^*} f_{W_t|W_{t-1},X_{t-1}^*} f_{X_{t-1}^*,W_{t-1},W_{t-2}} \tag{6.9}$$

which, in matrix notation (for fixed w_{t-1}), is

$$\begin{aligned} M_{W_t,w_{t-1},W_{t-2}} &= M_{W_t|w_{t-1},X_{t-1}^*} M_{X_{t-1}^*,w_{t-1},W_{t-2}} \\ \Rightarrow \quad M_{X_{t-1}^*,w_{t-1},W_{t-2}} &= M_{W_t|w_{t-1},X_{t-1}^*}^{-1} M_{W_t,w_{t-1},W_{t-2}} \end{aligned}$$

where $M_{W_t|w_{t-1},X_{t-1}^*}$ can be identified from the same procedure as in the identification of $M_{W_{t+1}|w_t,X_t^*}$. Hence, substituting the above into Eq. (6.8), we obtain the desired representation

$$\begin{aligned} & M_{w_t,X_t^*|w_{t-1},X_{t-1}^*} \left(M_{W_t|w_{t-1},X_{t-1}^*}^{-1} M_{W_t,w_{t-1},W_{t-2}} \right) = M_{W_{t+1}|w_t,X_t^*}^{-1} M_{W_{t+1},w_t,w_{t-1},W_{t-2}} \\ \Rightarrow \quad & M_{w_t,X_t^*|w_{t-1},X_{t-1}^*} = M_{W_{t+1}|w_t,X_t^*}^{-1} M_{W_{t+1},w_t,w_{t-1},W_{t-2}} \left(M_{W_t|w_{t-1},X_{t-1}^*}^{-1} M_{W_t,w_{t-1},W_{t-2}} \right)^{-1} \\ \Rightarrow \quad & M_{w_t,X_t^*|w_{t-1},X_{t-1}^*} = M_{W_{t+1}|w_t,X_t^*}^{-1} M_{W_{t+1},w_t,w_{t-1},W_{t-2}} M_{W_t,w_{t-1},W_{t-2}}^{-1} M_{W_t|w_{t-1},X_{t-1}^*}. \end{aligned} \tag{6.10}$$

Equation (6.10) implies that the Markov kernel $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$ for any fixed (w_t, w_{t-1}) as in matrix $M_{w_t,X_t^*|w_{t-1},X_{t-1}^*}$ can be identified from the observed distribution $f_{W_{t+1},W_t,W_{t-1},W_{t-2},W_{t-3}}$. Specifically, $f_{W_{t+1},W_t,W_{t-1},W_{t-2}}$ identifies $f_{W_{t+1}|W_t,X_t^*}$, $f_{W_t,W_{t-1},W_{t-2},W_{t-3}}$ identifies $f_{W_t|W_{t-1},X_{t-1}^*}$, and both $f_{W_{t+1}|W_t,X_t^*}$ and $f_{W_t|W_{t-1},X_{t-1}^*}$ leads to the identification of the Markov kernel $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$.

We summarize the identification in the discrete case as follows:

Assumption 6.1.2 Invertibility: *for any $w_t \in \mathcal{W}_t$, there exists $w_{t-1}, \overline{w}_t, \overline{w}_{t-1} \in \mathcal{W}_{t-1}$ such that,*

$$Rank \left(M_{W_{t-2},\widetilde{w}_{t-1},\widetilde{w}_t,W_{t+1}} \right) = K$$

for $(\widetilde{w}_{t-1}, \widetilde{w}_t)$ equal to (w_t, w_{t-1}), $(\overline{w}_t, w_{t-1})$, $(\overline{w}_t, \overline{w}_{t-1})$, and $(w_t, \overline{w}_{t-1})$; Furthermore,

$$k(w_t, \overline{w}_t, w_{t-1}, \overline{w}_{t-1}, \overline{x}_t^*) \neq k(w_t, \overline{w}_t, w_{t-1}, \overline{w}_{t-1}, \widetilde{x}_t^*)$$

for any $\overline{x}_t^* \neq \widetilde{x}_t^* \in \mathcal{X}_t^*$, where

$$k\left(w_t, \overline{w}_t, w_{t-1}, \overline{w}_{t-1}, x_t^*\right) = \frac{f_{W_t|W_{t-1},X_t^*}(w_t|w_{t-1}, x_t^*) f_{W_t|W_{t-1},X_t^*}(\overline{w}_t|\overline{w}_{t-1}, x_t^*)}{f_{W_t|W_{t-1},X_t^*}(\overline{w}_t|w_{t-1}, x_t^*) f_{W_t|W_{t-1},X_t^*}(w_t|\overline{w}_{t-1}, x_t^*)}.$$

This assumption is imposed on the observed distribution and is directly testable. In addi-

tion, notice that this assumption implies that for any $w_{t-1} \in \mathcal{W}_{t-1}$,

$$Rank\left(M_{W_{t-2},w_{t-1},W_t}\right) = K.$$

Specifically, $M_{W_{t-2},w_{t-1},W_t} = M^T_{W_t,w_{t-1},W_{t-2}}$ with

$$M_{W_t,w_{t-1},W_{t-2}} = M_{W_t|w_{t-1},X^*_{t-1}} M_{X^*_{t-1},w_{t-1},W_{t-2}}.$$

The invertibility of the two matrices on the right-hand side is implied by Assumption 6.1.2.

Assumption 6.1.3 Monotonicity and normalization: *For any $w_t \in \mathcal{W}_t$, One of the following conditions holds:*

1) $f_{W_{t+1}|W_t,X^*_t}\left(w_1|w_t,x^*_j\right) > f_{W_{t+1}|W_t,X^*_t}\left(w_1|w_t,x^*_{j+1}\right)$ for $j = 1,2,\ldots,K-1$;

2) The τ-th quantile of $f_{W_{t+1}|W_t,X^*_t}(\cdot|w_t,x^*)$ is monotonic in x^*.

3) There exists a function $\omega(\cdot)$ such that

$$E\left[\omega\left(W_{t+1}\right)|W_t = w_t, X^*_t = x^*_j\right] > E\left[\omega\left(W_{t+1}\right)|W_t = w_t, X^*_t = x^*_{j+1}\right].$$

In fact, condition 1) is a special case condition 3) with $\omega(x) = \delta(x = w_1)$.

Theorem 6.1.1 (Identification of Markov law of motion, discrete case): *Under the Assumptions 6.1.1, 6.1.2, and 6.1.3, the joint probability function $f_{W_{t+1},W_t,W_{t-1},W_{t-2},W_{t-3}}$ for any $t \in \{4,\ldots,T-1\}$ uniquely determines the Markov kernel $f_{W_t,X^*_t|W_{t-1},X^*_{t-1}}$.*

Notice that the equality

$$f_{W_t,W_{t-1}} = \sum_{X^*_{t-1}} f_{W_t|W_{t-1},X^*_{t-1}} f_{W_{t-1},X^*_{t-1}}$$

implies that, for any $w_{t-1} \in \mathcal{W}_t$,

$$\begin{aligned} &\overrightarrow{p}_{W_t,W_{t-1}=w_{t-1}} = M_{W_t|w_{t-1},X^*_{t-1}} \overrightarrow{p}_{W_{t-1}=w_{t-1},X^*_{t-1}} \\ \Leftrightarrow\ &\overrightarrow{p}_{W_{t-1}=w_{t-1},X^*_{t-1}} = M^{-1}_{W_t|w_{t-1},X^*_{t-1}} \overrightarrow{p}_{W_t,W_{t-1}=w_{t-1}} \end{aligned}$$

That means that the initial condition $f_{W_{t-1},X^*_{t-1}}$ is also identified.

Additionally, in the stationary case where $f_{W_{t+1}|W_t,X^*_t} = f_{W_t|W_{t-1},X^*_{t-1}}$, only four periods of data, i.e., $f_{W_{t+1},W_t,W_{t-1},W_{t-2}}$ are enough to identify the Makov kernel $f_{W_t,X^*_t|W_{t-1},X^*_{t-1}}$.

Implication of Assumptions in the Discrete Case

For the illustration purpose, we consider the case where W_t and X^*_t are both binary scalars: $\forall t,\ \text{supp}X^*_t = \text{supp}W_t \equiv \{0,1\}$. This is the simplest example of the models considered in our framework. One example of such a model is a binary version of Abbring, Chiappori, and Zavadil's 2008 "dynamic moral hazard" model of auto insurance. In that model, W_t is a binary indicator of claims occurrence, and X^*_t is a binary effort indicator, with $X^*_t = 1$ denoting higher effort. In this model, moral hazard in driving behavior and experience rating

in insurance pricing imply that the laws of motion for both W_t and X_t^* should exhibit state dependence:

$$\Pr(W_t = 1|w_{t-1}, x_t^*, x_{t-1}^*) = p(w_{t-1}, x_t^*); \quad \Pr(X_t^* = 1|x_{t-1}^*, w_{t-1}) = q(x_{t-1}^*, w_{t-1}). \quad (6.11)$$

These laws of motion satisfy Assumption 6.1.1.

Relative to the continuous case presented beforehand, some simplifications obtain in this finite-dimensional example. Notationally, the linear operators in the previous section reduce to matrices, with the L operators in the main proof corresponding to $K \times K$ square matrices, and the D operators to $K \times K$ diagonal matrices.

Assumptions 6.1.2 and 6.1.3 are quite transparent to interpret in the matrix setting. Assumption 6.1.2 implies the invertibility of certain matrices. As shown above, our identification results require that there exist at least four different points in the support of (W_t, W_{t-1}). In this dichotomous example, this implies that Assumptions 6.1.2 must hold for all four possible values of the pair (w_t, w_{t-1}). The following matrix equality holds, for all values of (w_t, w_{t-1}):

$$\begin{aligned} M_{W_{t+1},w_t,w_{t-1},W_{t-2}} &= M_{W_{t+1}|w_t,X_t^*} D_{w_t|w_{t-1},X_t^*} M_{X_t^*,w_{t-1},W_{t-2}} \quad (6.12) \\ &= M_{W_{t+1}|w_t,X_t^*} D_{w_t|w_{t-1},X_t^*} L_{X_t^*|w_{t-1},X_{t-1}^*} M_{X_{t-1}^*,w_{t-1},W_{t-2}}. \end{aligned}$$

Assumption 6.1.2 requires that the square matrix $M_{W_{t-2},w_{t-1},w_t,W_{t+1}} = M^T_{W_{t+1},w_t,w_{t-1},W_{t-2}}$ is invertible, which implies that $M_{W_{t+1},w_t,w_{t-1},W_{t-2}}$ is also invertible. This matrix is observed in the data, so that we can verify its invertibility directly.

Moreover, by Eq. (6.12), the invertibility of $M_{W_{t+1},w_t,w_{t-1},W_{t-2}}$ also implies the invertibility of $M_{W_{t+1}|w_t,X_t^*}$, $M_{X_t^*|w_{t-1},X_{t-1}^*}$, and $M_{X_{t-1}^*,w_{t-1},W_{t-2}}$, and that all the elements in the diagonal matrix $D_{w_t|w_{t-1},X_t^*}$ are nonzero.

Assumption 6.1.2 also puts restrictions on the eigenvalues in the spectral decomposition of the **AB** operator. In the discrete case, **AB** is an observed $K \times K$ matrix, and the spectral decomposition reduces to the usual matrix diagonalization. Assumption 6.1.6 implies that the eigenvalues are nonzero and finite, and that the eigenvalues are distinctive. For all (w_t, w_{t-1}), these assumptions can be verified, by directly diagonalizing the **AB** matrix.

In this discrete case, Assumption 6.1.3 is to an "ordering" assumption on the columns of the $M_{W_{t+1}|w_t,X_t^*}$ matrix, which are the eigenvectors of **AB**. This is because, for a matrix diagonalization $T = SDS^{-1}$, where D is diagonal, and T and S are square matrices, any permutation of the eigenvalues (the diagonal elements in D) and their corresponding eigenvectors (the columns in S) results in the same diagonal representation of T.

In order to compare values of X_t^* across these two periods, we must invoke Assumption 6.1.3 to pin down values of X_t^* which are consistent across the two periods. For this example, one reasonable monotonicity restriction is

$$\text{for } w_t = \{0, 1\}: \quad \mathbb{E}[W_{t+1}|w_t, X_t^* = 1] < \mathbb{E}[W_{t+1}|w_t, X_t^* = 0] \quad (6.13)$$

The restriction (6.13) implies that future claims W_{t+1} occur less frequently with higher

effort today, and imposes additional restrictions on the the $p(\cdots)$ and $q(\cdots)$ functions in (6.11).[4]

To see how this restriction orders the eigenvectors, and pins down the value of X_t^*, note that $\mathbb{E}[W_{t+1}|w_t, X_t^*] = f(W_{t+1} = 1|w_t, X_t^*)$, which is the second component of each eigenvector. Therefore, the monotonicity restriction (6.13) implies that the eigenvectors (and their corresponding eigenvalues) should be ordered such that their second components are decreasing, from left to right. Given this ordering, we assign a value of $X_t^* = 0$ to the eigenvector in the first column, and $X_t^* = 1$ to the eigenvector in the second column.

6.1.3 The Discrete Case versus a Finite Mixture Model

Here, we provide some additional comparison with the results in Kasahara and Shimotsu (2009) (KS), and show that KS's identification results are not applicable to the dynamic models with time-varying unobservables considered in Hu and Shum (2012).

We start by summarizing KS's main results. Throughout, we state KS's results using the notation in this paper. Since KS assume that the unobserved heterogeneity X^* is time-invariant, we attach no t subscript to it.[5] Using the notation in this paper, the second equality of KS's Eq. (3) is:

$$\begin{aligned} & f_{Y_1,M_1,\ldots,Y_T,M_T} \\ = & \sum_{X^*} f_{X^*} f_{M_1,Y_1|X^*} \prod_{t-2}^{T} f_{M_t|M_{t-1},Y_{t-1},\ldots,M_1,Y_1,X^*} f_{Y_t|M_t,M_{t-1},Y_{t-1},X^*}. \end{aligned}$$

In their baseline model (ie. their Assumption 1), they assume that the unobserved heterogeneity X^* does not affect the law of motion for the observed state variable M_t, and that Y_t is independent of (M_{t-1}, Y_{t-1}) conditional on M_t and X^*. This leads to

$$\frac{f_{Y_1,M_1,\ldots,Y_T,M_T}}{\prod_{t-2}^{T} f_{M_t|M_{t-1},Y_{t-1}}} = \sum_{X^*} f_{X^*} f_{M_1,Y_1|X^*} \prod_{t-2}^{T} f_{Y_t|M_t,X^*}, \qquad \text{[Eq. (9) in KS (2009)]}$$

which is Eq. (9) in KS. Notice that the LHS of the above is observed, and they demonstrate (in their Proposition 1) that the unknown densities on the RHS are identified from the observed quantity on the LHS for $T \geq 3$. In fact, such a setting with $T \geq 3$ forms a 3-measurement model as in section 2.5 so that its identification results based on Hu (2008) apply.

In section 3.2 of their paper, they consider a first-order Markovian model where the observed variables W_t can depend on W_{t-1} and X^*. They show that, by using $T \geq 6$

[4] See Hu (2008) for a number of other alternative ordering assumptions for the discrete case.

[5] The correspondence between KS's notation and Hu and Shum's is as follows:

$$\overbrace{\begin{bmatrix} a_t, x_t, s_t, m, \pi^m, Q^m(s_t|s_{t-1}) \\ P_t^m(a_t|x_t, x_{t-1}, a_{t-1}) \end{bmatrix}}^{\text{KS (2009)}} \Leftrightarrow \overbrace{\begin{bmatrix} Y_t, M_t, W_t, X^*, f_{X^*}, f_{W_t|W_{t-1},X^*} \\ f_{Y_t|M_t,M_{t-1},Y_{t-1},X^*} \end{bmatrix}}^{\text{our notation}}.$$

periods of data $W_1 \ldots, W_T$, and fixing the values in the odd periods $w_1, w_3, w_5, ..., w_{T-1}$, one obtains

$$f_{w_1,W_2,w_3,W_4,...,w_{T-1},W_T} = \sum_{X^*} f_{w_1,X^*} \left(\prod_{t=2,4,...}^{T-2} f_{w_{t+1},W_t|w_{t-1},X^*} \right) f_{W_T|w_{T-1},X^*}, \quad \text{[Eq. (27) in KS (2009)]}$$

which is Eq. (27) in KS. As they note, Eq. (27) has the same "independent marginals" form as Eq. (9), so that their identification scheme also applies to first-order Markov process with time-invariant X^* for $T \geq 6$. This is their Proposition 6.

However, this scheme no longer works in the case where the latent variable X_t^* varies over time, even if X_t^* is discrete. To see this, we consider a joint first-order Markov process $\{W_t, X_t^*\}$ where both W_t and X_t^* vary over time, as in Example 1 in the main text of this paper. Analogously to Eq. (27) in KS, we may have

$$f_{w_1,W_2,w_3,W_4,...,w_{T-1},W_T} = \sum_{X_{T-1}^*} ... \sum_{X_5^*} \sum_{X_3^*} \sum_{X_1^*} f_{w_1,X_1^*} \left(\prod_{t=2,4,...}^{T-2} f_{w_{t+1},X_{t+1}^*,W_t|w_{t-1},X_{t-1}^*} \right) f_{W_T|w_{T-1},X_{T-1}^*}.$$

Obviously, this takes a very different form than Eq. (27) above, because the components on the RHS involve values of the latent variable X_t^* in different periods. Hence, KS's identification scheme does not apply here. Notice that using more periods of data only exacerbates the problem; the more periods of data one uses, the more latent variables X_t^* appear when X_t^* is time-varying.

In conclusion, the identification strategy in KS does not apply to models where X_t^* is time-varying, even if X_t^* is discrete. An important innovation of the present paper is that it provides nonparametric identification for dynamic models with time-varying unobserved variables.

6.1.4 Assumptions in the Continuous Case

Consider a dynamic process $\{(W_T, X_T^*), \ldots, (W_t, X_t^*), ..., (W_1, X_1^*)\}_i$ for agent i. We assume that for each agent i, $\{(W_T, X_T^*), \ldots, (W_t, X_t^*), ..., (W_1, X_1^*)\}_i$ is an independent random draw from a bounded continuous distribution $f_{(W_T,X_T^*),...,(W_t,X_t^*),...,(W_1,X_1^*)}$. The researcher observes a panel dataset consisting of an i.i.d. random sample of $\{W_T, W_{T-1}, \ldots, W_1\}_i$, with $T \geq 5$, for many agents i. We first consider identification in the nonstationary case, where the Markov law of motion $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$ varies across periods. This model subsumes the special case of unobserved heterogeneity, in which X_t^* is fixed across all periods.

Next, we introduce our four assumptions. The first assumption below restricts attention to certain classes of models, while Assumptions 6.1.5-6.1.7 establish identification for the restricted class of models. Unless otherwise stated, all assumptions are taken to hold for all periods t.

Assumption 6.1.4 *(i)* First-order Markov: $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*,\Omega_{<t-1}} = f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$, *where* $\Omega_{<t-1} \equiv \{W_{t-2}, ..., W_1, X_{t-2}^*, ..., X_1^*\}$, *the history up to (but not including)* $t-1$. *(ii)* Limited feedback: $f_{W_t|W_{t-1},X_t^*,X_{t-1}^*} = f_{W_t|W_{t-1},X_t^*}$.

Assumption 6.1.4(i), a first-order Markov assumption, is satisfied for Markovian dynamic decision models (cf. Rust (1994)). Assumption 6.1.4(ii) is a "limited feedback" assumption, which rules out direct feedback from the last period's USV, X^*_{t-1}, on the current value of the observed W_t. When $W_t = (Y_t, M_t)$, as before, Assumption 1 implies:

$$\begin{aligned} f_{W_t|W_{t-1},X^*_t,X^*_{t-1}} &= f_{Y_t,M_t|Y_{t-1},M_{t-1},X^*_t,X^*_{t-1} \\ &= f_{Y_t|M_t,Y_{t-1},M_{t-1},X^*_t,X^*_{t-1}} \cdot f_{M_t|Y_{t-1},M_{t-1},X^*_t,X^*_{t-1}} \\ &= f_{Y_t|M_t,X^*_t,Y_{t-1},M_{t-1}} \cdot f_{M_t|Y_{t-1},M_{t-1},X^*_t}. \end{aligned}$$

In the bottom line of the above display, the limited feedback assumption eliminates X^*_{t-1} as a conditioning variable in both terms. In Markovian dynamic optimization models, the first term (the CCP) further simplifies to $f_{Y_t|M_t,X^*_t}$, because the Markovian laws of motion for (M_t, X^*_t) imply that the optimal policy function depends just on the current state variables. Hence, Assumption 1 imposes weaker restrictions on the first term than Markovian dynamic optimization models.[6]

In the second term of the above display, the limited feedback condition rules out direct feedback from last period's unobserved state variable X^*_{t-1} to the current observed state variable M_t. However, it allows indirect effects via X^*_{t-1}'s influence on Y_{t-1} or M_{t-1}. Implicitly, the limited feedback assumption 6.1.4(ii) imposes a timing restriction, that X^*_t is realized before M_t, so that M_t depends on X^*_t. While this is less restrictive than the assumption that M_t evolves independently of both X^*_{t-1} and X^*_t, which has been made in many applied settings to enable the estimation of the M_t law of motion directly from the data, it does rule out models such as $M_t = h(M_{t-1}, X^*_{t-1}) + \eta_t$, which implies the alternative timing assumption that X^*_t is realized after M_t.[7] For the special case of unobserved heterogeneity, where $X^*_t = X^*_{t-1}$, $\forall t$, the limited feedback assumption is trivial. Finally, the limited feedback assumption places no restrictions on the law of motion for X^*_t, and allows X^*_t to depend stochastically on $X^*_{t-1}, Y_{t-1}, M_{t-1}$. ■

For this paper, we assume that the unobserved state variable X^*_t is scalar-valued, and is drawn from a continuous distribution.[8] An important role in the identification argument is played by many integral equalities which demonstrate the equivalence of multivariate density functions which contain the latent variable X^*_t as an argument (which are not identified directly in the data), and those containing only observed variables W_t (which are identified directly from the data). To avoid cumbersome repetition, we will express these integral equalities in the convenient notation of linear operators, which we introduce here.

[6] Moreover, if we move outside the class of these models, the above display also shows that Assumption 1 does not rule out the dependence of Y_t on Y_{t-1} or M_{t-1}, which corresponds to some models of state dependence. These may include linear or nonlinear panel data models with lagged dependent variables, and serially correlated errors, cf. Arellano and Honore (2000). (Arellano, 2003, chs. 7–8) considers linear panel models with lagged dependent variables and serially-correlated unobservables, which is also related to our framework.

[7] Most empirical applications of dynamic optimization models with unobserved state variables satisfy the Markov and limited feedback conditions: examples from the industrial organization literature include Erdem et al. (2003), Crawford and Shum (2005), Das et al. (2007), Xu (2007), and Hendel and Nevo (2006).

[8] A discrete distribution for X^*_t, which is assumed in many applied settings (eg. Arcidiacono and Miller (2011)) is a special case, which we will consider as an example in Section 4.1.

Let R_1, R_2, R_3 denote three random variables, with support $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$, distributed with joint density $f_{R_1,R_2,R_3}(r_1,r_2,r_3)$ with support $\mathcal{R}_1 \times \mathcal{R}_2 \times \mathcal{R}_3$.[9] The linear operator L_{R_1,r_2,R_3} is a mapping from the $\mathcal{L}^p$-space of functions of R_3 to the $\mathcal{L}^p$ space of functions of R_1,[10] defined as [11]

$$\left(L_{R_1,r_2,R_3}h\right)(r_1) = \int f_{R_1,R_2,R_3}(r_1,r_2,r_3)h(r_3)dr_3; \quad h \in \mathcal{L}^p\left(\mathcal{R}_3\right),\ r_2 \in \mathcal{R}_2.$$

Similarly, we define the diagonal (or multiplication) operator

$$\left(D_{r_1|r_2,R_3}h\right)(r_3) = f_{R_1|R_2,R_3}(r_1|r_2,r_3)h(r_3); \quad h \in \mathcal{L}^p\left(\mathcal{R}_3\right),\ r_1 \in \mathcal{R}_1,\ r_2 \in \mathcal{R}_2.$$

In the next section, we show that our identification argument relies on a spectral decomposition of a linear operator generated from $L_{W_{t+1},w_t,w_{t-1},W_{t-2}}$, which corresponds to the observed density $f_{W_{t+1},W_t,W_{t-1},W_{t-2}}$. (A spectral decomposition is the operator analog of the eigenvalue-eigenvector decomposition for matrices, in the finite-dimensional case.)[12] The next two assumptions ensure the validity and uniqueness of this decomposition.

Assumption 6.1.5 Invertibility: *There exists variable(s) $V \subseteq W$ such that*
(i) for any $w_t \in \mathcal{W}_t$, there exists a $w_{t-1} \in \mathcal{W}_{t-1}$ and a neighborhood $\mathcal{N}^r$ around (w_t, w_{t-1})[13] *such that, for any $(\overline{w}_t, \overline{w}_{t-1}) \in \mathcal{N}^r$, $L_{V_{t-2},\overline{w}_{t-1},\overline{w}_t,V_{t+1}}$ is one-to-one;*
(ii) for any $w_t \in \mathcal{W}_t$, $L_{V_{t+1}|w_t,X_t^*}$ is one-to-one;
(iii) for any $w_{t-1} \in \mathcal{W}_{t-1}$, L_{V_{t-2},w_{t-1},V_t} is one-to-one.

Assumption 6.1.5 enables us to take inverses of certain operators, and is analogous to assumptions made in the nonclassical measurement error literature. Specifically, treating V_{t-2} and V_{t+1} as noisy "measurements" of the latent X_t^*, Assumption 6.1.5(i,ii) imposes the same restrictions between the measurements and the latent variable as (Hu and Schennach, 2008, Assumption 3) and (Carroll et al., 2010, Assumption 2.4). Compared with these two papers, Assumption 6.1.5(iii) is an extra assumption we need because, in our dynamic setting, there is a second latent variable, X_{t-1}^*, in the Markov law of motion $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$. Below, we show that Assumption 2(ii) implies that pre-multiplication by the inverse operator $L^{-1}_{V_{t+1}|w_t,X_t^*}$ is valid, while 2(i,iii) imply that post-multiplication by, respectively, $L^{-1}_{V_{t+1},w_t,w_{t-1},V_{t-2}}$ and $L^{-1}_{V_t,w_{t-1},V_{t-2}}$ is valid.[14]

[9] Here, capital letters denote random variables, while lower-case letters denote realizations.

[10] For $1 \leq p < \infty$, $\mathcal{L}^p(\mathcal{X})$ is the space of measurable real functions $h(\cdot)$ integrable in the L^p-norm, ie. $\int_{\mathcal{X}} |h(x)|^p d\mu(x) < \infty$, where μ is a measure on a σ-field in $\mathcal{X}$. One may also consider other classes of functions, such as bounded functions in $\mathcal{L}^1$, in the definition of an operator.

[11] Analogously, the operator $L_{R_1|r_2,R_3}$, corresponding to the conditional density $f_{R_1|R_2,R_3}$, is defined, for all functions $h \in \mathcal{L}^p(\mathcal{R}_3)$, and $r_2 \in \mathcal{R}_2$ as $\left(L_{R_1|r_2,R_3}h\right)(r_1) = \int f_{R_1|R_2,R_3}(r_1|r_2,r_3)h(r_3)dr_3$.

[12] Specifically, when W_t, X_t^* are both scalar and discrete with J ($< \infty$) points of support, the operator $L_{W_{t+1},w_t,w_{t-1},W_{t-2}}$ is a $J \times J$ matrix, and spectral decomposition reduces to diagonalization of the corresponding matrix. This discrete case is discussed in detail in Section 4.1.

[13] A neighborhood of $w \in \mathbb{R}^k$ is defined as $\left\{\overline{w} \in \mathbb{R}^k : \|\overline{w} - w\|_E < r\right\}$ for some $r > 0$, where $\|\cdot\|_E$ is the Euclidean metric.

[14] Additional details are given in Section 3 of the online appendix (Hu and Shum (2009)).

The statements in Assumption 6.1.5 are equivalent to *completeness* conditions which have recently been employed in the nonparametric IV literature: namely, an operator L_{R_1,r_2,R_3} is one-to-one if the corresponding density function f_{R_1,r_2,R_3} satisfies a "completeness" condition: for any r_2,

$$\left(L_{R_1,r_2,R_3}h\right)(r_1) = \int f(r_1,r_2,r_3)h(r_3)dr_3 = 0 \text{ for all } r_1 \text{ implies } h(r_3) = 0 \text{ for all } r_3. \tag{6.14}$$

Completeness is a high-level condition, and special cases of it have been considered in, eg. Newey and Powell (2003), Blundell et al. (2007a), D'Haultfoeuille (2011). However, sufficient conditions are not available for more general settings. Below, in Section 4, we will construct examples which satisfy the completeness requirements.

The variable(s) $V_t \subseteq W_t$ defined in Assumption 6.1.5 may be scalar, multidimensional, or W_t itself. Intuitively, by Assumption 6.1.5(ii), the variable(s) V_{t+1} are components of W_{t+1} which "transmit" information on the latent X_t^* conditional on W_t, the observables in the previous period. We consider suitable choices of V for specific examples in Section 4.[15] Assumption 6.1.5(ii) also rules out models where X_t^* has a continuous support, but W_{t+1} contains only discrete components. In this case, there is no subset $V_{t+1} \subseteq W_{t+1}$ for which $L_{V_{t+1}|w_t,X_t^*}$ can be one-to-one. Hence, dynamic discrete-choice models with a continuous unobserved state variable X_t^*, but only discrete observed state variables M_t, fail this assumption, and may be nonparametrically underidentified without further assumptions. Moreover, models where the W_t and X_t^* processes evolve independently will also fail this assumption. ■

Assumption 6.1.6 Uniqueness of spectral decomposition: *For any $w_t \in \mathcal{W}_t$ and any $\overline{x}_t^* \neq \widetilde{x}_t^* \in \mathcal{X}_t^*$, there exists a $w_{t-1} \in \mathcal{W}_{t-1}$ and corresponding neighborhood $\mathcal{N}^r$ satisfying Assumption 6.1.5(i) such that, for some $(\overline{w}_t, \overline{w}_{t-1}) \in \mathcal{N}^r$ with $\overline{w}_t \neq w_t$, $\overline{w}_{t-1} \neq w_{t-1}$:*
(i) $0 < k\left(w_t, \overline{w}_t, w_{t-1}, \overline{w}_{t-1}, x_t^*\right) < C < \infty$ for any $x_t^* \in \mathcal{X}_t^*$ and some constant C;
(ii) $k(w_t, \overline{w}_t, w_{t-1}, \overline{w}_{t-1}, \overline{x}_t^*) \neq k(w_t, \overline{w}_t, w_{t-1}, \overline{w}_{t-1}, \widetilde{x}_t^*)$, where

$$k\left(w_t, \overline{w}_t, w_{t-1}, \overline{w}_{t-1}, x_t^*\right) = \frac{f_{W_t|W_{t-1},X_t^*}(w_t|w_{t-1},x_t^*) f_{W_t|W_{t-1},X_t^*}(\overline{w}_t|\overline{w}_{t-1},x_t^*)}{f_{W_t|W_{t-1},X_t^*}(\overline{w}_t|w_{t-1},x_t^*) f_{W_t|W_{t-1},X_t^*}(w_t|\overline{w}_{t-1},x_t^*)}.$$

Assumption 6.1.6 ensures the uniqueness of the spectral decomposition of a linear operator generated from $L_{V_{t+1},w_t,w_{t-1},V_{t-2}}$. As Eq. (6.30) below shows, the $k(\cdots)$ function in the assumption corresponds to the eigenvalues in this decomposition, so that conditions (i) and (ii) guarantee that these eigenvalues are, respectively, bounded and distinct across all values of x_t^*. In turn, this ensures that the corresponding eigenfunctions are linearly independent, so that the spectral decomposition is unique.[16] ■

[15] There may be multiple choices of V which satisfy Assumption 6.1.5. In this case, the model may be overidentified, and it may be possible to do specification testing. We do not explore this possibility here.

[16] In the case where $W_t = (Y_t, M_t)$ and $f_{W_t|W_{t-1},X_t^*} = f_{Y_t|M_t,X_t^*} \cdot f_{M_t|Y_{t-1},M_{t-1},X_t^*}$, Assumption 6.1.6 simplifies further. Specifically, because the CCP term $f_{Y_t|M_t,X_t^*}$ does not contain W_{t-1}, Eq. (6.30) below implies that the CCP term cancels out in the expression of eigenvalues in the spectral decomposition, so that Assumption 6.1.6 imposes restrictions only on the second term $f_{M_t|Y_{t-1},M_{t-1},X_t^*}$. See additional discussion in Example 2 below.

Assumption 6.1.7 Monotonicity and normalization: *For any $w_t \in \mathcal{W}_t$, there exists a known functional G such that $G\left[f_{V_{t+1}|W_t,X_t^*}(\cdot|w_t,x_t^*)\right]$ is monotonic in x_t^*. We normalize $x_t^* = G\left[f_{V_{t+1}|W_t,X_t^*}(\cdot|w_t,x_t^*)\right]$.*

The eigenfunctions in the aforementioned spectral decomposition correspond to the densities $f_{V_{t+1}|W_t,X_t^*}(\cdot|w_t,x_t^*)$, for all values of x_t^*. Since X_t^* is unobserved, the eigenfunctions are only identified up to an arbitrary one-to-one transformation of X_t^*. To resolve this issue, we need additional restrictions deriving from the economic or stochastic structure of the model, to "pin down" the values of the unobserved X_t^* relative to the observed variables. In Assumption 6.1.7, this additional structure comes in the form of the functional G which, when applied to the family of densities $f_{V_{t+1}|W_t,X_t^*}(\cdot|w_t,x_t^*)$, is monotonic in x_t^*, given w_t. Given the monotonicity restriction, we can normalize X_t^* by setting, $x_t^* = G\left[f_{V_{t+1}|W_t,X_t^*}(\cdot|w_t,x_t^*)\right]$ without loss of generality.[17] The functional G, which may depend on the value of w_t, could be the mean, mode, median, or another quantile of $f_{V_{t+1}|W_t,X_t^*}$. ■

Assumptions 1-4 are the four main assumptions underlying our identification arguments. Of these four assumptions, all except Assumption 2(i,iii) involve densities not directly observed in the data, and are not directly testable in the continuous case.

6.1.5 Nonparametric Identification in the Continuous Case

We present our argument for the nonparametric identification of the Markov law of motion $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$ by way of several intermediate lemmas. The first two lemmas present convenient representations of the operators corresponding to the observed density $f_{V_{t+1},w_t,w_{t-1},V_{t-2}}$ and the Markov law of motion $f_{w_t,X_t^*|w_{t-1},X_{t-1}^*}$, for given values of $(w_t,w_{t-1}) \in \mathcal{W}_t \times \mathcal{W}_{t-1}$:

Lemma 6.1.1 (Representation of the observed density $f_{V_{t+1},w_t,w_{t-1},V_{t-2}}$): *For any $t \in \{3,\ldots,T-1\}$, Assumption 6.1.4 implies that, for any $(w_t,w_{t-1}) \in \mathcal{W}_t \times \mathcal{W}_{t-1}$,*

$$L_{V_{t+1},w_t,w_{t-1},V_{t-2}} = L_{V_{t+1}|w_t,X_t^*} D_{w_t|w_{t-1},X_t^*} L_{X_t^*,w_{t-1},V_{t-2}}. \tag{6.15}$$

Proof: (Lemma 6.1.1)

[17] To be clear, the monotonicity assumption here is a model restriction, and not without loss of generality; if it were false, our identification argument would not recover the correct CCP's and laws of motion for the underlying model. See Matzkin (2003) and Hu and Schennach (2008) for similar uses of monotonicity restrictions in the context of nonparametric identification problems.

By Assumption 6.1.4(i), the observed density $f_{W_{t+1},W_t,W_{t-1},W_{t-2}}$ equals

$$\begin{aligned}
&\int\int f_{W_{t+1},W_t,X_t^*,X_{t-1}^*,W_{t-1},W_{t-2}} dx_t^* dx_{t-1}^* \\
= &\int\int f_{W_{t+1}|W_t,W_{t-1},W_{t-2},X_t^*,X_{t-1}^*} f_{W_t,X_t^*|W_{t-1},W_{t-2},X_{t-1}^*} f_{X_{t-1}^*,W_{t-1},W_{t-2}} dx_t^* dx_{t-1}^* \\
= &\int\int f_{W_{t+1}|W_t,X_t^*} f_{W_t,X_t^*|W_{t-1},X_{t-1}^*} f_{X_{t-1}^*,W_{t-1},W_{t-2}} dx_t^* dx_{t-1}^* \\
= &\int\int f_{W_{t+1}|W_t,X_t^*} f_{W_t|W_{t-1},X_t^*,X_{t-1}^*} f_{X_t^*|W_{t-1},X_{t-1}^*} f_{X_{t-1}^*,W_{t-1},W_{t-2}} dx_t^* dx_{t-1}^* \\
= &\int\int f_{W_{t+1}|W_t,X_t^*} f_{W_t|W_{t-1},X_t^*,X_{t-1}^*} f_{X_t^*|W_{t-1},W_{t-2},X_{t-1}^*} f_{X_{t-1}^*,W_{t-1},W_{t-2}} dx_t^* dx_{t-1}^* \\
= &\int\int f_{W_{t+1}|W_t,X_t^*} f_{W_t|W_{t-1},X_t^*,X_{t-1}^*} f_{X_t^*,X_{t-1}^*,W_{t-1},W_{t-2}} dx_t^* dx_{t-1}^*.
\end{aligned}$$

(We omit all the arguments in the density functions.) Assumption 6.1.4(ii) then implies

$$\begin{aligned}
f_{W_{t+1},W_t,W_{t-1},W_{t-2}} &= \int f_{W_{t+1}|W_t,X_t^*} f_{W_t|W_{t-1},X_t^*} \left(\int f_{X_t^*,X_{t-1}^*,W_{t-1},W_{t-2}} dx_{t-1}^* \right) dx_t^* \\
&= \int f_{W_{t+1}|W_t,X_t^*} f_{W_t|W_{t-1},X_t^*} f_{X_t^*,W_{t-1},W_{t-2}} dx_t^*. \qquad (6.16)
\end{aligned}$$

In operator notation, given values of $(w_t, w_{t-1}) \in \mathcal{W}_t \times \mathcal{W}_{t-1}$, this is

$$L_{W_{t+1},w_t,w_{t-1},W_{t-2}} = L_{W_{t+1}|w_t,X_t^*} D_{w_t|w_{t-1},X_t^*} L_{X_t^*,w_{t-1},W_{t-2}}. \qquad (6.17)$$

For the variable(s) $V_t \subseteq W_t$, for all periods t, introduced in Assumption 6.1.5, Eq. (6.17) implies that the joint density of $\{V_{t+1}, W_t, W_{t-1}, V_{t-2}\}$ is expressed in operator notation as $L_{V_{t+1},w_t,w_{t-1},V_{t-2}} = L_{V_{t+1}|w_t,X_t^*} D_{w_t|w_{t-1},X_t^*} L_{X_t^*,w_{t-1},V_{t-2}}$, as postulated by Lemma 1. *Q.E.D.*

Lemma 6.1.2 (Representation of Markov law of motion): *For any $t \in \{3, \ldots, T-1\}$, Assumptions 6.1.4, 6.1.5(ii), and 6.1.5(iii) imply that, for any $(w_t, w_{t-1}) \in \mathcal{W}_t \times \mathcal{W}_{t-1}$,*

$$L_{w_t,X_t^*|w_{t-1},X_{t-1}^*} = L_{V_{t+1}|w_t,X_t^*}^{-1} L_{V_{t+1},w_t,w_{t-1},V_{t-2}} L_{V_t,w_{t-1},V_{t-2}}^{-1} L_{V_t|w_{t-1},X_{t-1}^*}. \qquad (6.18)$$

Proof: (Lemma 6.1.2)
Assumption 6.1.4 implies the following two equalities:

$$\begin{aligned}
f_{V_{t+1},W_t,W_{t-1},V_{t-2}} &= \int f_{V_{t+1}|W_t,X_t^*} f_{W_t,X_t^*,W_{t-1},V_{t-2}} dx_t^* \\
f_{W_t,X_t^*,W_{t-1},V_{t-2}} &= \int f_{W_t,X_t^*|W_{t-1},X_{t-1}^*} f_{X_{t-1}^*,W_{t-1},V_{t-2}} dx_{t-1}^*. \qquad (6.19)
\end{aligned}$$

In operator notation, for fixed w_t, w_{t-1}, the above equations are expressed:

$$\begin{aligned}
L_{V_{t+1},w_t,w_{t-1},V_{t-2}} &= L_{V_{t+1}|w_t,X_t^*} L_{w_t,X_t^*,w_{t-1},V_{t-2}} \\
L_{w_t,X_t^*,w_{t-1},V_{t-2}} &= L_{w_t,X_t^*|w_{t-1},X_{t-1}^*} L_{X_{t-1}^*,w_{t-1},V_{t-2}}.
\end{aligned}$$

Substituting the second line into the first, we get

$$\begin{aligned} L_{V_{t+1},w_t,w_{t-1},V_{t-2}} &= L_{V_{t+1}|w_t,X_t^*} L_{w_t,X_t^*|w_{t-1},X_{t-1}^*} L_{X_{t-1}^*,w_{t-1},V_{t-2}} \\ \Leftrightarrow L_{w_t,X_t^*|w_{t-1},X_{t-1}^*} L_{X_{t-1}^*,w_{t-1},V_{t-2}} &= L^{-1}_{V_{t+1}|w_t,X_t^*} L_{V_{t+1},w_t,w_{t-1},V_{t-2}}. \end{aligned} \tag{6.20}$$

where the second line uses Assumption 2(ii). Next, we eliminate $L_{X_{t-1}^*,w_{t-1},V_{t-2}}$ from the above. Again using Assumption 1, we have

$$f_{V_t,W_{t-1},V_{t-2}} = \int f_{V_t|W_{t-1},X_{t-1}^*} f_{X_{t-1}^*,W_{t-1},V_{t-2}} dx_{t-1}^* \tag{6.21}$$

which, in operator notation (for fixed w_{t-1}), is

$$L_{V_t,w_{t-1},V_{t-2}} = L_{V_t|w_{t-1},X_{t-1}^*} L_{X_{t-1}^*,w_{t-1},V_{t-2}} \quad \Rightarrow \quad L_{X_{t-1}^*,w_{t-1},V_{t-2}} = L^{-1}_{V_t|w_{t-1},X_{t-1}^*} L_{V_t,w_{t-1},V_{t-2}}$$

where the right-hand side applies Assumption 6.1.5(ii). Hence, substituting the above into Eq. (6.20), we obtain the desired representation

$$\begin{aligned} & L_{w_t,X_t^*|w_{t-1},X_{t-1}^*} L^{-1}_{V_t|w_{t-1},X_{t-1}^*} L_{V_t,w_{t-1},V_{t-2}} = L^{-1}_{V_{t+1}|w_t,X_t^*} L_{V_{t+1},w_t,w_{t-1},V_{t-2}} \\ \Rightarrow \quad & L_{w_t,X_t^*|w_{t-1},X_{t-1}^*} L^{-1}_{V_t|w_{t-1},X_{t-1}^*} = L^{-1}_{V_{t+1}|w_t,X_t^*} L_{V_{t+1},w_t,w_{t-1},V_{t-2}} L^{-1}_{V_t,w_{t-1},V_{t-2}} \\ \Rightarrow \quad & L_{w_t,X_t^*|w_{t-1},X_{t-1}^*} = L^{-1}_{V_{t+1}|w_t,X_t^*} L_{V_{t+1},w_t,w_{t-1},V_{t-2}} L^{-1}_{V_t,w_{t-1},V_{t-2}} L_{V_t|w_{t-1},X_{t-1}^*}. \end{aligned} \tag{6.22}$$

The second line applies Assumption 2(iii) to postmultiply by $L^{-1}_{V_t,w_{t-1},V_{t-2}}$, while in the third line, we postmultiply both sides by $L_{V_t|w_{t-1},X_{t-1}^*}$. *Q.E.D.*

Since $L_{V_{t+1},w_t,w_{t-1},V_{t-2}}$ and $L_{V_t,w_{t-1},V_{t-2}}$ are observed, Lemma 6.1.2 implies that the identification of the operators $L_{V_{t+1}|w_t,X_t^*}$ and $L_{V_t|w_{t-1},X_{t-1}^*}$ implies the identification of $L_{w_t,X_t^*|w_{t-1},X_{t-1}^*}$, the operator corresponding to the Markov law of motion. The next lemma postulates that $L_{V_{t+1}|w_t,X_t^*}$ is identified just from observed data.

Lemma 6.1.3 (Identification of $f_{V_{t+1}|W_t,X_t^*}$): *For any $t \in \{3,\ldots,T-1\}$, Assumptions 6.1.4, 6.1.5, 6.1.6, 6.1.7 imply that the density $f_{V_{t+1},W_t,W_{t-1},V_{t-2}}$ uniquely determines the density $f_{V_{t+1}|W_t,X_t^*}$.*

This lemma encapsulates the heart of the identification argument, which is the identification of $f_{V_{t+1}|W_t,X_t^*}$ via a spectral decomposition of an operator generated from the observed density $f_{V_{t+1},W_t,W_{t-1},V_{t-2}}$. Once this is established, re-applying Lemma 6.1.3 to the operator corresponding to the observed density $f_{V_t,W_{t-1},W_{t-2},V_{t-3}}$ yields the identification of $f_{V_t|W_{t-1},X_{t-1}^*}$. Once $f_{V_{t+1}|W_t,X_t^*}$ and $f_{V_t|W_{t-1},X_{t-1}^*}$ are identified, then so is the Markov law of motion $f_{w_t,X_t^*|w_{t-1},X_{t-1}^*}$, from Lemma 6.1.2.

Proof: (Lemma 6.1.3)

For each w_t, choose a w_{t-1} and a neighborhood $\mathcal{N}^r$ around (w_t, w_{t-1}) to satisfy Assumptions 6.1.5(i) and 6.1.6, and pick a $(\overline{w}_t, \overline{w}_{t-1})$ within the neighborhood $\mathcal{N}^r$ to satisfy Assumption 6.1.6. Because $(\overline{w}_t, \overline{w}_{t-1}) \in \mathcal{N}^r$, also $(\overline{w}_t, w_{t-1})$, $(\overline{w}_t, w_{t-1}) \in \mathcal{N}^r$. By Lemma 6.1.1, $L_{V_{t+1},w_t,w_{t-1},V_{t-2}} = L_{V_{t+1}|w_t,X_t^*} D_{w_t|w_{t-1},X_t^*} L_{X_t^*,w_{t-1},V_{t-2}}$. The first term on the RHS,

$L_{V_{t+1}|w_t,X_t^*}$, does not depend on w_{t-1}, and the last term $L_{X_t^*,w_{t-1},V_{t-2}}$ does not depend on w_t. This feature suggests that, by evaluating Eq. (6.15) at the four pairs of points (w_t, w_{t-1}), $(\overline{w}_t, w_{t-1})$, $(w_t, \overline{w}_{t-1})$, $(\overline{w}_t, \overline{w}_{t-1})$, each pair of equations will share one operator in common. Specifically:

$$\begin{aligned}
\text{for } (w_t, w_{t-1}) \quad &: \quad L_{V_{t+1},w_t,w_{t-1},V_{t-2}} = L_{V_{t+1}|w_t,X_t^*} D_{w_t|w_{t-1},X_t^*} L_{X_t^*,w_{t-1},V_{t-2}}, \quad (6.23)\\
\text{for } (\overline{w}_t, w_{t-1}) \quad &: \quad L_{V_{t+1},\overline{w}_t,w_{t-1},V_{t-2}} = L_{V_{t+1}|\overline{w}_t,X_t^*} D_{\overline{w}_t|w_{t-1},X_t^*} L_{X_t^*,w_{t-1},V_{t-2}}, \quad (6.24)\\
\text{for } (w_t, \overline{w}_{t-1}) \quad &: \quad L_{V_{t+1},w_t,\overline{w}_{t-1},V_{t-2}} = L_{V_{t+1}|w_t,X_t^*} D_{w_t|\overline{w}_{t-1},X_t^*} L_{X_t^*,\overline{w}_{t-1},V_{t-2}}, \quad (6.25)\\
\text{for } (\overline{w}_t, \overline{w}_{t-1}) \quad &: \quad L_{V_{t+1},\overline{w}_t,\overline{w}_{t-1},V_{t-2}} = L_{V_{t+1}|\overline{w}_t,X_t^*} D_{\overline{w}_t|\overline{w}_{t-1},X_t^*} L_{X_t^*,\overline{w}_{t-1},V_{t-2}}. \quad (6.26)
\end{aligned}$$

Assumption 6.1.5(ii) implies that $L_{V_{t+1}|\overline{w}_t,X_t^*}$ is invertible. Moreover, Assumption 6.1.6(i) implies $f_{W_t|W_{t-1},X_t^*}(\overline{w}_t|w_{t-1},x_t^*) > 0$ for all x_t^* so that $D_{\overline{w}_t|w_{t-1},X_t^*}$ is invertible. We can then solve for $L_{X_t^*,w_{t-1},V_{t-2}}$ from Eq. (6.24) as

$$D^{-1}_{\overline{w}_t|w_{t-1},X_t^*} L^{-1}_{V_{t+1}|\overline{w}_t,X_t^*} L_{V_{t+1},\overline{w}_t,w_{t-1},V_{t-2}} = L_{X_t^*,w_{t-1},V_{t-2}}.$$

Plugging in this expression to Eq. (6.23) leads to

$$L_{V_{t+1},w_t,w_{t-1},V_{t-2}} = L_{V_{t+1}|w_t,X_t^*} D_{w_t|w_{t-1},X_t^*} D^{-1}_{\overline{w}_t|w_{t-1},X_t^*} L^{-1}_{V_{t+1}|\overline{w}_t,X_t^*} L_{V_{t+1},\overline{w}_t,w_{t-1},V_{t-2}}.$$

Lemma 1 of Hu and Schennach (2008) shows that, given Assumption 6.1.5(i), we can post-multiply by $L^{-1}_{V_{t+1},\overline{w}_t,w_{t-1},V_{t-2}}$, to obtain:

$$\begin{aligned}
\mathbf{A} &\equiv L_{V_{t+1},w_t,w_{t-1},V_{t-2}} L^{-1}_{V_{t+1},\overline{w}_t,w_{t-1},V_{t-2}}\\
&= L_{V_{t+1}|w_t,X_t^*} D_{w_t|w_{t-1},X_t^*} D^{-1}_{\overline{w}_t|w_{t-1},X_t^*} L^{-1}_{V_{t+1}|\overline{w}_t,X_t^*}. \quad (6.27)
\end{aligned}$$

Similar manipulations of Eqs. (6.25) and Eq. (6.26) lead to

$$\begin{aligned}
\mathbf{B} &\equiv L_{V_{t+1},\overline{w}_t,\overline{w}_{t-1},V_{t-2}} L^{-1}_{V_{t+1},w_t,\overline{w}_{t-1},V_{t-2}}\\
&= L_{V_{t+1}|\overline{w}_t,X_t^*} D_{\overline{w}_t|\overline{w}_{t-1},X_t^*} D^{-1}_{w_t|\overline{w}_{t-1},X_t^*} L^{-1}_{V_{t+1}|w_t,X_t^*}. \quad (6.28)
\end{aligned}$$

Assumption 6.1.5(i) guarantees that, for any w_t, ($\overline{w}_t$, w_{t-1}, $\overline{w}_{t-1}$) exist so that (9) and (10) are valid operations. Finally, we postmultiply Eq. (6.27) by Eq. (6.28) to obtain

$$\begin{aligned}
\mathbf{AB} &= L_{V_{t+1}|w_t,X_t^*} D_{w_t|w_{t-1},X_t^*} D^{-1}_{\overline{w}_t|w_{t-1},X_t^*} \left(L^{-1}_{V_{t+1}|\overline{w}_t,X_t^*} L_{V_{t+1}|\overline{w}_t,X_t^*} \right) \times\\
&\quad \times D_{\overline{w}_t|\overline{w}_{t-1},X_t^*} D^{-1}_{w_t|\overline{w}_{t-1},X_t^*} L^{-1}_{V_{t+1}|w_t,X_t^*}\\
&= L_{V_{t+1}|w_t,X_t^*} \left(D_{w_t|w_{t-1},X_t^*} D^{-1}_{\overline{w}_t|w_{t-1},X_t^*} D_{\overline{w}_t|\overline{w}_{t-1},X_t^*} D^{-1}_{w_t|\overline{w}_{t-1},X_t^*} \right) L^{-1}_{V_{t+1}|w_t,X_t^*}\\
&\equiv L_{V_{t+1}|w_t,X_t^*} D_{w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},X_t^*} L^{-1}_{V_{t+1}|w_t,X_t^*}, \qquad \text{where} \quad (6.29)
\end{aligned}$$

$$\begin{aligned}\left(D_{w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},X_t^*}h\right)(x_t^*) &= \left(D_{w_t|w_{t-1},X_t^*}D^{-1}_{\overline{w}_t|w_{t-1},X_t^*}D_{\overline{w}_t|\overline{w}_{t-1},X_t^*}D^{-1}_{w_t|\overline{w}_{t-1},X_t^*}h\right)(x_t^*)\\ &= \frac{f_{W_t|W_{t-1},X_t^*}(w_t|w_{t-1},x_t^*)f_{W_t|W_{t-1},X_t^*}(\overline{w}_t|\overline{w}_{t-1},x_t^*)}{f_{W_t|W_{t-1},X_t^*}(\overline{w}_t|w_{t-1},x_t^*)f_{W_t|W_{t-1},X_t^*}(w_t|\overline{w}_{t-1},x_t^*)}h(x_t^*)\\ &\equiv k\left(w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},x_t^*\right)h\left(x_t^*\right).\end{aligned} \tag{6.30}$$

This equation implies that the observed operator **AB** on the left hand side of Eq. (6.29) has an inherent eigenvalue-eigenfunction decomposition, with the eigenvalues corresponding to the function $k\left(w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},x_t^*\right)$ and the eigenfunctions corresponding to the density $f_{V_{t+1}|W_t,X_t^*}(\cdot|w_t,x_t^*)$. The decomposition in Eq. (6.29) is similar to the decomposition in Hu and Schennach (2008) or Carroll et al. (2010).

Assumption 6.1.6 ensures that this decomposition is unique. Specifically, Eq. (6.29) implies that the operator **AB** on the LHS has the same spectrum as the diagonal operator $D_{w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},X_t^*}$. Assumption 6.1.6(i) guarantees that the spectrum of the diagonal operator $D_{w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},X_t^*}$ is bounded. Since an operator is bounded by the largest element of its spectrum, Assumption 6.1.6(i) also implies that the operator **AB** is bounded, whence we can apply Theorem XV.4.3.5 from Dunford and Schwartz (1971) to show the uniqueness of the spectral decomposition of bounded linear operators.

Several ambiguities remain in the spectral decomposition. First, Eq. (6.29) itself does not imply that the eigenvalues $k\left(w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},x_t^*\right)$ are distinctive for different values x_t^*. When the eigenvalues are the same for multiple values of x_t^*, the corresponding eigenfunctions are only determined up to an arbitrary linear combination, implying that they are not identified. Assumption 6.1.6(ii) rules out this possibility, and implies that for each w_t, we can find values $\overline{w}_t$, w_{t-1}, and $\overline{w}_{t-1}$ such that the eigenvalues are distinct across all x_t^*.[18,19]

Second, the eigenfunctions $f_{V_{t+1}|W_t,X_t^*}(\cdot|w_t,x_t^*)$ in the spectral decomposition (6.29) are unique up to multiplication by a scalar constant. However, these are density functions, so their scale is pinned down because they must integrate to one. Finally, both the eigenvalues and eigenfunctions are indexed by X_t^*. Since our arguments are nonparametric, and X_t^* is unobserved, we need an additional monotonicity condition, in Assumption 4, to pin down the value of X_t^* relative of the observed variables. This was discussed earlier, in the remarks following Assumption 4.

[18]Specifically, the operators **AB** corresponding to different values of $(\overline{w}_t,w_{t-1},\overline{w}_{t-1})$ share the same eigenfunctions $f_{V_{t+1}|W_t,X_t^*}(\cdot|w_t,x_t^*)$. Assumption 6.1.6(ii) implies that, for any two different eigenfunctions $f_{V_{t+1}|W_t,X_t^*}(\cdot|w_t,x_t^*)$ and $f_{V_{t+1}|W_t,X_t^*}(\cdot|w_t,\widetilde{x}_t^*)$, one can always find values of $(\overline{w}_t,w_{t-1},\overline{w}_{t-1})$ such that the two different eigenfunctions correspond to two different eigenvalues, i.e., $k\left(w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},x_t^*\right)\neq k\left(w_t,\overline{w}_t,w_{t-1},\overline{w}_{t-1},\widetilde{x}_t^*\right)$.

[19]When w_t (resp. w_{t-1}) is close to $\overline{w}_t$ (resp. $\overline{w}_{t-1}$), Eq. (6.30) implies that the logarithm of the eigenvalues in this decomposition can be represented as a second-order derivative of the log-density $f_{W_t|W_{t-1},X_t^*}$. Therefore, a sufficient condition for 6.1.6(ii) is that $\frac{\partial^3}{\partial z_t\partial z_{t-1}\partial x_t^*}\log f_{W_t|W_{t-1},X_t^*}$ is continuous and nonzero, which implies that $\frac{\partial^2}{\partial z_t\partial z_{t-1}}\log f_{W_t|W_{t-1},X_t^*}$ is monotonic in x_t^* for any (w_t,w_{t-1}), where z_t is the continuous component of w_t.

Therefore, altogether the density $f_{V_{t+1}|W_t,X_t^*}$ or $L_{V_{t+1}|w_t,X_t^*}$ is nonparametrically identified for any given $w_t \in \mathcal{W}_t$ via the spectral decomposition in Eq. (6.29). *Q.E.D.*

By re-applying Lemma 6.1.3 to the observed density $f_{V_t,W_{t-1},W_{t-2},V_{t-3}}$, it follows that the density $f_{V_t|W_{t-1},X_{t-1}^*}$ is identified.[20] Hence, by Lemma 6.1.2, we have shown the following result:

Theorem 6.1.2 (Identification of Markov law of motion, non-stationary case): *Under the Assumptions 6.1.4, 6.1.5, 6.1.6, and 6.1.7, the density* $f_{W_{t+1},W_t,W_{t-1},W_{t-2},W_{t-3}}$ *for any* $t \in \{4,\ldots,T-1\}$ *uniquely determines the density* $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$.

Initial Conditions

Some CCP-based estimation methodologies for dynamic optimization models (eg. Hotz et al. (1994), Bajari et al. (2007a)) require simulation of the Markov process $(W_t, X_t^*, W_{t+1}, X_{t+1}^*, W_{t+2}, X_{t+2}^*, \ldots)$ starting from some initial values W_{t-1}, X_{t-1}^*. When there are unobserved state variables, this raises difficulties because X_{t-1}^* is not observed. However, it turns out that, as a by-product of the main identification results, we are also able to identify the marginal densities f_{W_{t-1},X_{t-1}^*}. For any given initial value of the observed variables w_{t-1}, knowledge of f_{W_{t-1},X_{t-1}^*} allows us to draw an initial value of X_{t-1}^* consistent with w_{t-1}.

Corollary 6.1.1 (Identification of initial conditions, non-stationary case): *Under the Assumptions 6.1.4, 6.1.5, 6.1.6, and 6.1.7, the density* $f_{W_{t+1},W_t,W_{t-1},W_{t-2},W_{t-3}}$ *for any* $t \in \{4,\ldots,T-1\}$ *uniquely determines the density* f_{W_{t-1},X_{t-1}^*}.

Proof: (Corollary 6.1.1)

From Lemma 6.1.3, $f_{V_t|W_{t-1},X_{t-1}^*}$ is identified from density $f_{V_t,W_{t-1},W_{t-2},V_{t-3}}$. The equality $f_{V_t,W_{t-1}} = \int f_{V_t|W_{t-1},X_{t-1}^*} f_{W_{t-1},X_{t-1}^*} dx_{t-1}^*$ implies that, for any $w_{t-1} \in \mathcal{W}_t$,

$$\begin{aligned} & f_{V_t,W_{t-1}=w_{t-1}} = L_{V_t|w_{t-1},X_{t-1}^*} f_{W_{t-1}=w_{t-1},X_{t-1}^*} \\ \Leftrightarrow\ & f_{W_{t-1}=w_{t-1},X_{t-1}^*} = L^{-1}_{V_t|w_{t-1},X_{t-1}^*} f_{V_t,W_{t-1}=w_{t-1}} \end{aligned}$$

where the second line applies Assumption 6.1.5(ii). Hence, $f_{W_{t-1},X_{t-1}^*}$is identified. *Q.E.D.*

Stationarity

In the proof of Theorem 6.1.2 from the previous section, we only use the fifth period of data W_{t-3} for the identification of $L_{V_t|w_{t-1},X_{t-1}^*}$. Given that we identify $L_{V_{t+1}|w_t,X_t^*}$ using four periods of data, i.e., $\{W_{t+1}, W_t, W_{t-1}, W_{t-2}\}$, the fifth period W_{t-3} is not needed when $L_{V_t|w_{t-1},X_{t-1}^*} = L_{V_{t+1}|w_t,X_t^*}$. This is true when the Markov kernel density $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$ is time-invariant. Thus, in the stationary case, only four periods of data, $\{W_{t+1}, W_t, W_{t-1}, W_{t-2}\}$, are required to identify $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$. Formally, we make the additional assumption:

[20] Recall that Assumptions 1-4 are assumed to hold for all periods t. Hence, applying Lemma 6.1.3 to the observed density $f_{V_t,W_{t-1},W_{t-2},V_{t-3}}$ does not require any additional assumptions.

Assumption 6.1.8 Stationarity: *the Markov law of motion of* (W_t, X_t^*) *is time-invariant:* $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*} = f_{W_2,X_2^*|W_1,X_1^*}$, $\forall\ 2 \leq t \leq T$.

Stationarity is usually maintained in infinite-horizon dynamic programming models. Given the foregoing discussion, we present the next corollary without proof.

Corollary 6.1.2 (Identification of Markov law of motion, stationary case): *Under assumptions 6.1.4, 6.1.5, 6.1.6, 6.1.7, and 6.1.8, the observed density* $f_{W_{t+1},W_t,W_{t-1},W_{t-2}}$ *for any* $t \in \{3,\ldots,T-1\}$ *uniquely determines the density* $f_{W_2,X_2^*|W_1,X_1^*}$.

In the stationary case, initial conditions are still a concern. The following corollary, analogous to Corollary 6.1.1 for the non-stationary case, postulates the identification of the marginal density f_{W_t,X_t^*}, for periods $t \in \{1,\ldots,T-3\}$. For any of these periods, f_{W_t,X_t^*} can be used as a sampling density for the initial conditions.[21]

Corollary 6.1.3 (Identification of initial conditions, stationary case): *Under assumptions 6.1.4, 6.1.5, 6.1.6, 6.1.7, and 6.1.8, the observed density* $f_{W_{t+1},W_t,W_{t-1},W_{t-2}}$ *for any* $t \in \{3,\ldots,T-1\}$ *uniquely determines the density* f_{W_{t-2},X_{t-2}^*}.

Proof: (Corollary 6.1.3)

Under stationarity, the operator $L_{V_{t-1}|w_{t-2},X_{t-2}^*}$ is the same as $L_{V_{t+1}|w_t,X_t^*}$, which is identified from the observed density $f_{V_{t+1},W_t,W_{t-1},V_{t-2}}$ (by Lemma 6.1.3). Because $f_{V_{t-1},W_{t-2}} = \int f_{V_{t-1}|W_{t-2},X_{t-2}^*} f_{W_{t-2},X_{t-2}^*} dx_{t-2}^*$, the same argument as in the proof of Corollary 6.1.1 then implies that f_{W_{t-2},X_{t-2}^*} is identified from the observed density $f_{V_{t-1},W_{t-2}}$. *Q.E.D.*

6.1.6 Comments on Assumptions in Specific Examples

Even though we focus on nonparametric identification, we believe that our results can be valuable for applied researchers working in a parametric setting, because they provide a guide for specifying models such that they are nonparametrically identified. As part of a pre-estimation check, our identification assumptions could be verified for a prospective model via direct calculation, as in the examples here. If the prospective model satisfies the assumptions, then the researcher could proceed to estimation, with the confidence that underlying variation in the data, rather than the particular functional forms chosen, is identifying the model parameters. If some assumptions are violated, then our results suggest ways that the model could be adjusted in order to be nonparametrically identified.

To this end, we present an example of dynamic models here. Because some of the assumptions that we made for our identification argument are quite abstract, we discuss these assumptions in the context of these examples.[22]

[21] Even in the stationary case, where $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$ is invariant over time, the marginal density of f_{W_{t-1},X_{t-1}^*} may still vary over time (unless the Markov process (W_t, X_t^*) starts from the steady-state). For this reason, it is useful to identify f_{W_t,X_t^*} across a range of periods.

[22] A third example, based on Rust (1987), is in the supplemental material (Hu and Shum (2009)).

An Example: Generalized Investment Model

For the second example, we consider a dynamic model of firm R&D and product quality in the "generalized dynamic investment" framework described in Doraszelski and Pakes (2007).[23] In this model, $W_t = (Y_t, M_t)$, where Y_t is a firm's R&D in year t, and M_t is the product's installed base. The unobserved state variable X_t^* is the firm's product quality, which is unobserved by the econometrician but observed by the firm, and affects their R&D choices.

Product quality $X_t^* \in \mathbb{R}$ evolves as follows:

$$X_t^* = 0.8X_{t-1}^* + 0.2\exp\left(\psi\left(Y_{t-1}\right)\right)\nu_t. \tag{6.31}$$

In the above, $\nu_t \in \mathbb{R}$ is a standard normal shock, distributed independently over t, and $\psi(\cdot) < \infty$, $\psi'(\cdot) > 0$. Eq. (6.31) implies $f_{X_t^*|Y_{t-1},M_{t-1},X_{t-1}^*} = f_{X_t^*|Y_{t-1},X_{t-1}^*}$.

Installed base evolves as:

$$M_{t+1} = M_t[1 + \exp(\eta_{t+1} + X_{t+1}^*)] \tag{6.32}$$

where $\eta_{t+1} \in \mathbb{R}$ is a random shock following the extreme value distribution, with density $f_{\eta_{t+1}}(\eta) = \exp(\eta - e^{\eta})$ for $\eta \in \mathbb{R}$, independently across t. This law of motion also implies that $f_{M_{t+1}|Y_t,M_t,X_t^*,X_{t+1}^*} = f_{M_{t+1}|M_t,X_{t+1}^*}$. Eq. (6.32) implies that, *ceteris paribus*, product quality raises installed base. Moreover, we also assume that the initial installed base $M_1 > 0$, so that $M_t > 0$ for all t and, for a given M_t, $M_{t+1} \in (M_t, +\infty)$.

Each period, a firm chooses its R&D to maximize its discounted future profits:

$$\begin{aligned} Y_t &= Y^*(M_t, X_t^*, \gamma_t) \\ &= \text{argmax}_{0\le y\le \bar{I}}\big[\underbrace{\Pi(M_t, X_t^*)}_{\text{profits}} - \underbrace{\gamma_t}_{\text{shock}} \cdot \underbrace{Y_t^2}_{\text{R\&D cost}} + \beta\mathbb{E}\underbrace{V(M_{t+1}, X_{t+1}^*, \gamma_{t+1})}_{\text{value fxn}}\big] \end{aligned} \tag{6.33}$$

$\bar{I}$ is a cap on per-period R&D, and γ_t is a shock to R&D costs. We assume that $\gamma_t \in (0, +\infty)$ follows a standard exponential distribution independently across t. The RHS of Eq. (6.33) is supermodular in Y_t and $-\gamma_t$, for all (M_t, X_t^*); accordingly, for fixed (M_t, X_t^*), the firm's optimal R&D investment Y_t^* is monotonically decreasing in γ_t, and take values in $(0, \bar{I}]$.

We verify the assumptions out of order, leaving the most involved Assumption 2 to the end. Since we focus here on the stationary case, without loss of generality we label the four observed periods of data W_t as $t = 1, 2, 3, 4$.

Assumption 1 is satisfied for this model. **Assumption 6.1.6** contains two restrictions on the density $f_{W_3|W_2,X_3^*}$, which factors as

$$f_{W_3|W_2,X_3^*} = f_{Y_3|M_3,X_3^*} \cdot f_{M_3|M_2,X_3^*}. \tag{6.34}$$

The first term in Eq. (6.34) is the density of R&D Y_3. Because the first term is not a function of M_2, Eq. (6.30) implies that the investment density $f_{Y_3|M_3,X_3^*}$ cancels out from

[23] See (Hu and Shum, 2009, Section 1.2) for additional discussion of dynamic investment models.

the numerator and denominator of the eigenvalues in the spectral decomposition as follows:

$$\begin{aligned} k\left(w_3, \overline{w}_3, w_2, \overline{w}_2, x_3^*\right) &= \frac{f_{W_3|W_2,X_3^*}(w_3|w_2,x_3^*) f_{W_3|W_2,X_3^*}(\overline{w}_3|\overline{w}_2,x_3^*)}{f_{W_3|W_2,X_3^*}(\overline{w}_3|w_2,x_3^*) f_{W_3|W_2,X_3^*}(w_3|\overline{w}_2,x_3^*)} \\ &= \frac{f_{M_3|M_2,X_3^*}(m_3|m_2,x_3^*) f_{M_3|M_2,X_3^*}(\overline{m}_3|\overline{m}_2,x_3^*)}{f_{M_3|M_2,X_3^*}(\overline{m}_3|m_2,x_3^*) f_{M_3|M_2,X_3^*}(m_3|\overline{m}_2,x_3^*)}. \end{aligned} \tag{6.35}$$

Hence, to ensure that the eigenvalues are distinct, we only require $f_{Y_3|M_3,X_3^*} > 0$ for all X_3^*. Given the discussions above, conditional on (M_3, X_3^*), investment Y_3 will be monotonically decreasing in the shock γ_3. Since, by assumption, the density of γ_3 is nonzero for $\gamma_3 > 0$, so also the conditional density $f_{Y_3|M_3,X_3^*} > 0$ along its support $(0, \bar{I}]$, for all (M_3, X_3^*), as required.

The second term $f_{M_3|M_2,X_3^*}$ is the law of motion for installed base which, by assumption, is an extreme value distribution with density

$$\begin{aligned} f_{M_3|M_2,X_3^*}(m_3|m_2,x_3^*) &= \frac{1}{(m_3-m_2)} \exp\left[\log\left(\frac{m_3-m_2}{m_2}\right) - x_3^* - e^{\log\left(\frac{m_3-m_2}{m_2}\right)-x_3^*}\right] \\ &= \frac{e^{-x_3^*}}{m_2} \exp\left(-e^{-x_3^*}\left[\frac{m_3-m_2}{m_2}\right]\right). \end{aligned}$$

Plugging this into Eq. (6.35), we obtain an expression for the eigenvalues

$$k\left(w_3, \overline{w}_3, w_2, \overline{w}_2, x_3^*\right) = \exp\left(-e^{-x_3^*}\left[\frac{-\left(\overline{m}_3 - m_3\right)\left(\overline{m}_2 - m_2\right)}{m_2 \overline{m}_2}\right]\right). \tag{6.36}$$

For given m_3, we can pick a finite and nonzero m_2,[24] and set $(\overline{m}_3, \overline{m}_2) = (m_3 - \Delta, m_2 + \Delta)$, with Δ nonzero and small. At these values, the eigenvalues in Eq. (6.36) simplify to $\exp\left(-e^{-x_3^*}\left[\frac{\Delta^2}{m_2(m_2+\Delta)}\right]\right)$ so that, for fixed m_3, and $x_3^* \in \mathbb{R}$, $0 < k\left(w_3, \overline{w}_3, w_2, \overline{w}_2, x_3^*\right) < 1$, which satisfies Assumption 6.1.6(i). Moreover, the eigenvalues in Eq. (6.36) are monotonic in x_3^* for any given $(w_3, \overline{w}_3, w_2, \overline{w}_2)$, which implies Assumption 6.1.6(ii).

To verify **Assumption 6.1.7**, we set $V_t = M_t$ for all t. Note $\mathbb{E}[\log \frac{M_4 - m_3}{m_3} | m_3, y_3, x_3^*] = \mathbb{E}[\eta_4] + \mathbb{E}[X_4^* | x_3^*, y_3]$. Because the law of motion for product quality $X_4^* = 0.8 X_3^* + 0.2 \exp\left(\psi\left(Y_3\right)\right) \nu_4$ implies that $\mathbb{E}[X_4^* | x_3^*, y_3]$ is monotonic in x_3^*, we set the functional G to be $x_3^* = \mathbb{E}[\log \frac{M_4 - m_3}{m_3} | m_3, y_3, x_3^*]$.

Finally, **Assumption 6.1.5** contains three injectivity assumptions. As before, we use $V_t = M_t$, for all periods t. Here, we provide sufficient conditions for Assumption 2, in the context of this investment model. We exploit the fact that the laws of motion for this model (cf. Eqs. (6.31) and (6.32)) are either linear or log-linear to apply results from the convolution literature, for which operator invertibility has been studied in detail.

For Assumption 2, it is sufficient to establish the injectivity of the operators L_{M_1,w_2,w_3,M_4}, $L_{M_4|w_3,X_3^*}$, and L_{M_1,w_2,M_3} for any (w_2, w_3) in the support. We start by showing the injectivity of L_{M_4,w_3,w_2,M_1}, $L_{M_4|w_3,X_3^*}$, and L_{M_3,w_2,M_1}. As shown in the proof of Lemma 1,

[24] In verifying Assumption 2(i) below, we show that the assumption holds for all (w_3, w_2), so that the neighborhood $\mathcal{N}^r$ is unrestricted. Hence, in verifying Assumption 3(i) here, we can pick any m_2, and also pick any other point $(\overline{m}_3, \overline{m}_2)$ as needed.

Assumption 1 implies that

$$\begin{aligned} L_{M_4,w_3,w_2,M_1} &= L_{M_4|w_3,X_3^*} D_{w_3|w_2,X_3^*} L_{X_3^*,w_2,M_1} \\ &= L_{M_4|w_3,X_3^*} D_{w_3|w_2,X_3^*} L_{X_3^*|w_2,X_2^*} L_{X_2^*,w_2,M_1} \qquad (6.37) \\ L_{M_3,w_2,M_1} &= L_{M_3|w_2,X_2^*} L_{X_2^*,w_2,M_1}. \qquad (6.38) \end{aligned}$$

Furthermore, we have $L_{M_4|w_3,X_3^*} = L_{M_4|w_3,X_4^*} L_{X_4^*|w_3,X_3^*}$.

Hence, the injectivity of L_{M_4,w_3,w_2,M_1}, $L_{M_4|w_3,X_3^*}$, and L_{M_3,w_2,M_1} is implied by the injectivity of $L_{M_4|w_3,X_4^*}$, $D_{w_3|w_2,X_3^*}$, $L_{X_3^*|w_2,X_2^*}$ and $L_{X_2^*,w_2,M_1}$.[25] It turns out that assumptions we have made already for this example ensure that three of these operators are injective. We discuss each case in turn.

(i) The diagonal operator $D_{w_3|w_2,X_3^*}$ has kernel function $f_{w_3|w_2,X_3^*} = f_{y_3|m_3,X_3^*} f_{m_3|m_2,X_3^*}$. In the discussion on Assumption 6.1.6(i) above, we showed that $f_{y_3|m_3,X_3^*}$ is nonzero along its support and that $f_{m_3|m_2,X_3^*}$ is nonzero for any (m_3, m_2, x_3^*) in the support. Therefore, $D_{w_3|w_2,X_3^*}$ is injective.

(ii) For $L_{M_4|w_3,X_4^*}$, we use Eq. (6.32) whereby, for every (y_3, m_3), M_4 is a convolution of X_4^*, ie. $\log[M_4 - M_3] - \log M_3 = X_4^* + \eta_4$. We have

$$\begin{aligned} g(m_4) &\equiv \left(L_{M_4|w_3,X_4^*} h\right)(m_4) \\ &= \int_{-\infty}^{\infty} f_{M_4|w_3,X_4^*}(m_4|w_3, x_4^*)\, h(x_4^*) dx_4^* \\ &= \int_{-\infty}^{\infty} \frac{1}{m_4 - m_3} f_{\eta_4}\left(\log\left(\frac{m_4 - m_3}{m_3}\right) - x_4^*\right) h(x_4^*) dx_4^* \\ &= \frac{1}{m_4 - m_3} \int_{-\infty}^{\infty} f_{\eta_4}(\varphi_4 - x_4^*)\, h(x_4^*) dx_4^*, \qquad \left[\varphi_4 \equiv \log\left(\frac{m_4 - m_3}{m_3}\right)\right] \\ &\equiv \frac{1}{m_4 - m_3} \times \left(L_{\varphi_4|X_4^*} h\right)(\varphi_4) \end{aligned}$$

Since the function $\frac{1}{m_4 - m_3}$ is nonzero, $g(m_4) = 0$ for any $m_4 \in (m_3, \infty)$ implies $\left(L_{\varphi_4|X_4^*} h\right)(\varphi_4) = 0$ for any $\varphi_4 \in \mathbb{R}$, where the kernel of the operator $L_{\varphi_4|X_4^*}$ has a convolution form $f_{\eta_4}(\varphi_4 - x_4^*)$. As shown in Lemma 6.1.4, as long as the characteristic function of η_4 has no real zeros, which is satisfied by the assumed extreme value distribution,[26] the corresponding operator $L_{\varphi_4|X_4^*}$ is injective. Therefore, $\left(L_{\varphi_4|X_4^*} h\right)(\varphi_4) = 0$ for any $\varphi_4 \in \mathbb{R}$ implies $h(x_4^*) = 0$ for any $x_4^* \in \mathbb{R}$. Thus, the operator $L_{M_4|w_3,X_4^*}$ is injective.

(iii) Similarly, for fixed w_2, X_3^* is a convolution of X_2^*, ie. $X_3^* = 0.8X_2^* + 0.2\exp(\psi(Y_2))\,\nu_3$ (cf. Eq. (6.31)). By an argument similar to that for the previous operator, we can show that $L_{X_3^*|w_2,X_2^*}$ is injective.

(iv) For the last operator, corresponding to the density $f_{X_2^*,w_2,M_1}$, the model assumptions do not allow us to establish injectivity directly. This is because this joint density confounds

[25] By stationarity, the operators $L_{M_4|w_3,X_3^*}$ and $L_{M_3|w_2,X_2^*}$ are the same, and do not need to be considered separately. Our notion of stationarity here is distinct from the notion of covariance-stationarity for stochastic processes. Indeed, as defined in Eq. (6.32), the M_t process may not be covariance-stationary, but the law of motion $f_{M_4|w_3,X_4^*}$ is still time-invariant.

[26] The characteristic function for η_4 is $\phi_{\eta_4}(\tau) = \Gamma(1 + i\tau)$, which is nonzero for any $\tau \in \mathbb{R}$.

both the structural components (laws of motion) in the model and the initial condition $f_{X_1^*,M_1}$. Thus in general, injectivity of this operator is not verifiable based only on the assumptions made thus far about the laws of motion for the state variables.

However, in the special case where product quality X_t^* evolves exogenously – that is, $\psi(\cdot) = 0$ in Eq. (6.31) – it turns out that an additional independence assumption on the initial values of the state variables (X_1^*, M_1), i.e., $f_{X_1^*,M_1} = f_{X_1^*} f_{M_1}$, suffices to ensure injectivity of the operator $L_{X_2^*,w_2,M_1}$:

Claim 1: If $\psi(\cdot) = 0$ in Eq. (6.31), and the initial values of the state variables (X_1^*, M_1) are independently distributed, the operator $L_{X_2^*,w_2,M_1}$ is injective.
Proof: in Appendix B.

Up to this point, we have shown the injectivity of L_{M_4,w_3,w_2,M_1}, $L_{M_4|w_3,X_3^*}$, and L_{M_3,w_2,M_1}. It turns out that this implies injectivity of L_{M_1,w_2,w_3,M_4} and L_{M_1,w_2,M_3}, as required by Assumption 6.1.5:

Claim 2: L_{M_1,w_2,w_3,M_4} and L_{M_1,w_2,M_3} are injective.
Proof: in Appendix B.

The assumptions underlying Claim 1, particularly the assumption that X_t^* evolves exogenously, are restrictive. However, we stress here that these are sufficient conditions, and are not necessary for the general results. Moreover, a large class of investment models (eg. Olley and Pakes (1996), Levinsohn and Petrin (2000)) assume that the unobserved variable X_t^* (denoting productivity) evolves exogenously. Finally, these assumptions are needed only in this example because we assume X_t^* to be continuous-valued. As Example 1 above demonstrates, when X_t^* is discrete, we can verify the identification assumptions even when the evolution of X_t^* depends on past values of the observed variables w_{t-1}.

6.1.7 Summary

We have considered the identification of a first-order Markov process $\{W_t, X_t^*\}$ when only $\{W_t\}$ is observed. Under non-stationarity, the Markov law of motion $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$ is identified from the distribution of the five observations $W_{t+1}, \ldots, W_{t-3}$. Under stationarity, identification of $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$ obtains with only four observations $W_{t+1}, \ldots, W_{t-2}$. Once $f_{W_t,X_t^*|W_{t-1},X_{t-1}^*}$ is identified, nonparametric identification of the remaining parts of the models – particularly, the per-period utility functions – can proceed by applying the results in Magnac and Thesmar (2002) and Bajari et al. (2007b), who considered dynamic models without unobserved state variables X_t^*.

For a general k-th order Markov process ($k < \infty$), it can be shown that the $3k+2$ observations $W_{t+k}, \ldots, W_{t-2k-1}$ can identify the Markov law of motion $f_{W_t,X_t^*|W_{t-1},\ldots,W_{t-k},X_{t-1}^*,\ldots,X_{t-k}^*}$, under appropriate extensions of the assumptions in this paper.

We have only considered the case where the unobserved state variable X_t^* is scalar-valued. The case where X_t^* is a multivariate process, which may apply to dynamic game

settings, presents some serious challenges. Specifically, when X_t^* is multi-dimensional, Assumption 6.1.5(ii), which requires that $L_{V_{t+1}|w_t,X_t^*}$ be one-to-one, can be quite restrictive. (Ackerberg et al., 2007, Section 2.4.3) discuss the difficulties with multivariate unobserved state variables in the context of dynamic investment models.

Finally, this paper has focused on identification, but not estimation. In ongoing work, we are using our identification results to guide the estimation of dynamic models with unobserved state variables. This would complement recent papers on the estimation of parametric dynamic models with unobserved state variables, using non-CCP-based approaches.[27]

6.1.8 Proofs

Proofs of Claims for Example 2

Here we provide the proofs for Claims 1 and 2 in example 2. We start with a general lemma regarding integral operators based on a convolution form, which is useful for what follows. We consider the basic convolution case where $X = Z + \epsilon$ with $Z \in \mathbb{R}$, $\epsilon \in \mathbb{R}$, and $Z \perp \epsilon$. The independence between Z and ϵ implies that $f_{X|Z}(x|z) = f_\epsilon(x - z)$. We define the two operators

$$\begin{aligned} \left(L_{X|Z}h\right)(x) &= \int f_\epsilon(x-z)\,h(z)\,dz \\ \left(L_{X|Z}^*h\right)(z) &= \int f_\epsilon(x-z)\,h(x)\,dx. \end{aligned} \tag{6.39}$$

Notice that $L_{X|Z}^*$ maps functions of X to those of Z.

Lemma 6.1.4 *Suppose that (i) the kernel of operator $L_{X|Z}$ is $f_\epsilon(x-z)$; (ii) the Fourier transform of f_ϵ does not vanish on the real line. Then, operators $L_{X|Z}$ and $L_{X|Z}^*$ are injective.*

Proof: (Lemma 6.1.4)
We have

$$\begin{aligned} g(x) &\equiv \left(L_{X|Z}h\right)(x) \\ &= \int f_\epsilon(x-z)\,h(z)\,dz. \end{aligned}$$

Let ϕ_g denote the Fourier transform of g, and ϕ_ϵ that of f_ϵ. We have for any $t \in \mathbb{R}$

$$\phi_g(t) = \phi_\epsilon(t)\phi_h(t).$$

Therefore, $\phi_g = 0$ implies $\phi_h = 0$ if $\phi_\epsilon(t) \neq 0$ for any $t \in \mathbb{R}$, which is assumed by hypothesis. So $L_{X|Z}$ is injective.

[27] Imai et al. (2009) and Norets (2009) consider Bayesian estimation, and Fernandez-Villaverde and Rubio-Ramirez (2007) consider efficient simulation estimation based on particle filtering.

Next, we show the injectivity of $L^*_{X|Z}$. We consider

$$\begin{aligned}\varphi(z) &\equiv \left(L^*_{X|Z}\psi\right)(z)\\ &= \int f_\epsilon(x-z)\,\psi(x)\,dx\\ &\equiv \int \kappa(z-x)\,\psi(x)dx\end{aligned}$$

where $\kappa(x) \equiv f_\epsilon(-x)$, i.e., $\phi_\kappa(t) = \phi_\epsilon(-t)$. We then have

$$\begin{aligned}\phi_\varphi(t) &= \phi_\kappa(t)\phi_\psi(t)\\ &= \phi_\epsilon(-t)\phi_\psi(t).\end{aligned}$$

Again, $\phi_\varphi = 0$ implies $\phi_\psi = 0$ because $\phi_\epsilon(t) \neq 0$ for any $t \in \mathbb{R}$. Thus, $L^*_{X|Z}$ is injective. *Q.E.D.*

Given this lemma, we proceed to prove the two claims from Example 2.

Proof of Claim 1: The operator $L_{X_2^*,w_2,M_1}$ has kernel function

$$\begin{aligned}f_{X_2^*,w_2,M_1} &= \int\int f_{X_2^*,y_2,m_2,X_1^*,Y_1,M_1}\,dy_1 dx_1^*\\ &= f_{y_2|m_2,X_2^*} f_{m_2|X_2^*,M_1} \int\int f_{X_2^*|Y_1,X_1^*} f_{Y_1|X_1^*,M_1} f_{X_1^*,M_1}\,dy_1 dx_1^*\\ &= f_{y_2|m_2,X_2^*} f_{m_2|X_2^*,M_1} \int f_{X_2^*|X_1^*}\left(\int f_{Y_1|X_1^*,M_1}\,dy_1\right) f_{X_1^*,M_1}\,dx_1^*\\ &= f_{y_2|m_2,X_2^*} f_{m_2|X_2^*,M_1}\left(\int f_{X_2^*|X_1^*} f_{X_1^*}\,dx_1^*\right) f_{M_1}\\ &= f_{y_2|m_2,X_2^*} f_{X_2^*} f_{m_2|X_2^*,M_1} f_{M_1}\end{aligned}$$

In the third line, we have utilized the restriction that $\psi(\cdot) = 0$ in Eq. (6.31) so that the density of $f_{Y_1|X_1^*,M_1}$ can be integrated out. The fourth line applies the independence of (X_1^*, M_1) so that $f_{X_1^*,M_1} = f_{X_1^*} f_{M_1}$. The corresponding operator equation is

$$L_{X_2^*,w_2,M_1} = D_{y_2|m_2,X_2^*} D_{X_2^*} L_{m_2|X_2^*,M_1} D_{M_1}. \tag{6.40}$$

Given that all the densities in the diagonal operators are nonzero and bounded, it remains

to show the injectivity of $L_{m_2|X_2^*,M_1}$. For a fixed m_2, we have:

$$\begin{aligned}
g\left(x_2^*\right) &\equiv \left(L_{m_2|X_2^*,M_1}h\right)\left(x_2^*\right) \\
&= \int_0^{m_2} f_{m_2|X_2^*,M_1}\left(m_2|x_2^*,m_1\right)h(m_1)dm_1 \\
&= \int_0^{m_2} \frac{1}{m_2-m_1} f_{\eta_2}\left(\log\left(\frac{m_2-m_1}{m_1}\right)-x_2^*\right)h(m_1)dm_1 \\
&= \int_0^{m_2} \frac{1}{m_2-m_1}\left(\frac{-m_2}{(m_2-m_1)\,m_1}\right)^{-1} f_{\eta_2}\left(\log\left(\frac{m_2-m_1}{m_1}\right)-x_2^*\right)h(m_1)d\log\left(\frac{m_2-m_1}{m_1}\right) \\
&= \int_{m_2}^{0} \frac{m_1}{m_2} f_{\eta_2}\left(\log\left(\frac{m_2-m_1}{m_1}\right)-x_2^*\right)h(m_1)d\log\left(\frac{m_2-m_1}{m_1}\right) \\
&= \int_{-\infty}^{\infty} f_{\eta_2}\left(\varphi_2-x_2^*\right)h\left(\frac{m_2}{e^{\varphi_2}+1}\right)\frac{1}{e^{\varphi_2}+1}d\varphi_2, \qquad \left[\varphi_2 \equiv \log\left(\frac{m_2-m_1}{m_1}\right)\right] \\
&\equiv \int_{-\infty}^{\infty} f_{\eta_2}\left(\varphi_2-x_2^*\right)\widetilde{h}\left(\varphi_2\right)d\varphi_2, \qquad \left[\widetilde{h}\left(\varphi_2\right) \equiv h\left(\frac{m_2}{e^{\varphi_2}+1}\right)\frac{1}{e^{\varphi_2}+1}\right] \\
&= \left(L^*_{\varphi_2|X_2^*}\widetilde{h}\right)\left(x_2^*\right),
\end{aligned}$$

where the operator $L^*_{\varphi_2|X_2^*}$ is defined analogously to Eq. (6.39). As shown above, $g\left(x_2^*\right) = 0$ for any $x_2^* \in \mathbb{R}$ implies that $\left(L^*_{\varphi_2|X_2^*}\widetilde{h}\right)\left(x_2^*\right) = 0$ for any $x_2^* \in \mathbb{R}$, where the kernel of $L^*_{\varphi_2|X_2^*}$ has a convolution form $f_{\eta_2}\left(\varphi_2 - x_2^*\right)$. Since the characteristic function of η_2 has no zeros on the real line, we can apply Lemma 4 to obtain the injectivity of $L^*_{\varphi_2|X_2^*}$. Accordingly, $\left(L^*_{\varphi_2|X_2^*}\widetilde{h}\right)\left(x_2^*\right) = 0$ for any $x_2^* \in \mathbb{R}$ implies $\widetilde{h}\left(\varphi_2\right) = 0$ for any $\varphi_2 \in \mathbb{R}$. Next, because $\widetilde{h}\left(\varphi_2\right) = h\left(\frac{m_2}{e^{\varphi_2}+1}\right)\frac{1}{e^{\varphi_2}+1}$ and $\frac{1}{e^{\varphi_2}+1}$ is nonzero, $\widetilde{h}\left(\varphi_2\right) = 0$ for any $\varphi_2 \in \mathbb{R}$ implies $h\left(\frac{m_2}{e^{\varphi_2}+1}\right) = 0$ for any $\varphi_2 \in \mathbb{R}$. Given $\varphi_2 \equiv \log\left(\frac{m_2-m_1}{m_1}\right)$, we have $h\left(m_1\right) = 0$ for any $m_1 \in (0, m_2)$. Altogether, then, $g\left(x_2^*\right) = 0$ for any $x_2^* \in \mathbb{R}$ implies $h\left(m_1\right) = 0$ for any $m_1 \in (0, m_2)$, thus demonstrating the injectivity of the operator $L_{m_2|X_2^*,M_1}$, as claimed. *Q.E.D.*

Proof of Claim 2: First, we show the injectivity of L_{M_1,w_2,w_3,M_4}. For fixed (w_2, w_3):

$$\begin{aligned}
f_{M_1,w_2,w_3,M_4} &= \int f_{M_4|w_3,X_3^*} f_{w_3|w_2,X_3^*} f_{X_3^*,w_2,M_1} dx_3^*. \\
&= \int \left(\int f_{M_4|w_3,X_4^*} f_{X_4^*|w_3,X_3^*} dx_4^*\right) f_{w_3|w_2,X_3^*} \left(\int f_{X_3^*|w_2,X_2^*} f_{X_2^*,w_2,M_1} dx_2^*\right) dx_3^* \\
&= \int \left(\int f_{M_4|w_3,X_4^*} f_{X_4^*|w_3,X_3^*} dx_4^*\right) f_{w_3|w_2,X_3^*} \left(\int f_{X_3^*|w_2,X_2^*} f_{y_2|m_2,X_2^*} f_{X_2^*} f_{m_2|X_2^*,M_1} f_{M_1} dx_2^*\right) dx_3^*.
\end{aligned}$$

Therefore, the equivalent operator equation is

$$\begin{aligned}
L_{M_1,w_2,w_3,M_4} &= L_{M_1,y_2,m_2,y_3,m_3,M_4} \\
&= D_{M_1} L^*_{m_2|X_2^*,M_1} D_{X_2^*} D_{y_2|m_2,X_2^*} L^*_{X_3^*|w_2,X_2^*} D_{w_3|w_2,X_3^*} L^*_{X_4^*|w_3,X_3^*} L^*_{M_4|w_3,X_4^*} \qquad (6.41)
\end{aligned}$$

In the above, the L^* operators are defined analogously to Eq. (6.39), and all the L^* operators are based on convolution kernels. Earlier, in the main text and Claim 1, we showed that the

operators $L_{m_2|X_2^*,M_1}$, $L_{X_3^*|w_2,X_2^*}$, $L_{X_4^*|w_3,X_3^*}$, and $L_{M_4|w_3,X_4^*}$ are injective; hence, by applying Lemma 4, we also obtain the injectivity of $L^*_{m_2|X_2^*,M_1}$, $L^*_{X_3^*|w_2,X_2^*}$, $L^*_{X_4^*|w_3,X_3^*}$, and $L^*_{M_4|w_3,X_4^*}$ using an argument similar to that used in the proof of Claim 1 above.

Finally, all the densities corresponding to the diagonal operators in Eq.(6.41) are nonzero and bounded, implying that these operators are injective. Hence, L_{M_1,w_2,w_3,M_4} is also injective.

Second, for L_{M_1,w_2,M_3}, we have

$$\begin{aligned} f_{M_1,w_2,M_3} &= \int f_{M_3|w_2,X_2^*} f_{X_2^*,w_2,M_1} dx_2^* \\ &= \int \left(\int f_{M_3|w_2,X_3^*} f_{X_3^*|w_2,X_2^*} dx_3^* \right) f_{X_2^*,w_2,M_1} dx_2^* \\ &= \int \left(\int f_{M_3|w_2,X_3^*} f_{X_3^*|w_2,X_2^*} dx_3^* \right) f_{y_2|m_2,X_2^*} f_{X_2^*} f_{m_2|X_2^*,M_1} f_{M_1} dx_2^*. \end{aligned}$$

Therefore, the equivalent operator equation is

$$L_{M_1,w_2,M_3} = D_{M_1} L^*_{m_2|X_2^*,M_1} D_{X_2^*} D_{y_2|m_2,X_2^*} L^*_{X_3^*|w_2,X_2^*} L^*_{M_3|w_2,X_3^*}.$$

By stationarity, the injectivity of $L^*_{M_3|w_2,X_3^*}$ is implied by that of $L^*_{M_4|w_3,X_4^*}$. All the other operators on the RHS also appeared in Eq. (6.41), and we argued above that these were injective. Thus, L_{M_1,w_2,M_3} is injective. *Q.E.D.*

6.2 Closed-Form Estimation of DDC with Unobserved State Variables

Although the nonparametric identification is quite general, it is still useful for empirical research to provide a relatively simple estimator for a particular specification of the model as long as such a specification can capture the key economic causality in the model. Given the difficulty in the estimation of dynamic discrete choice models with unobserved state variables, Hu and Sasaki (2018) consider a popular parametric specification of the model and provide a closed-form estimator for the inputs of the conditional choice probability estimator. Let d_t denote firms' exit decisions based on their productivity x_t^* and other covariates w_t. The law of motion of the productivity is

$$x_t^* = \alpha^d + \gamma^d x_{t-1}^* + \eta_t^d \text{ if } d_{t-1} = d \in \{0, 1\}. \tag{6.42}$$

In addition, they use residuals from the production function as a proxy x_t for latent x_t^* satisfying

$$x_t = x_t^* + \epsilon_t. \tag{6.43}$$

Therefore, they obtain

$$x_{t+1} = \alpha^d + \gamma^d x_t^* + \eta_{t+1}^d + \epsilon_{t+1} \tag{6.44}$$

Under the assumption that the error terms η_t^d and ϵ_t are random shocks, they first estimate the coefficients $\left(\alpha^d, \gamma^d\right)$ using other covariates M_t as instruments. The distribution of the error term ϵ_t can then be estimated using Kotlarski's identity. Furthermore, they are able to provide a closed-form expression for the conditional choice probability $\Pr\left(d_t|x_t^*, w_t\right)$ as a function of observed distribution functions.

6.2.1 Background

Forward-looking agents making dynamic decisions based on unobserved state variables are of interest in economic researches. While econometricians may not observe the true state variables, they often have access to or can construct proxy variables. To estimate the dynamic discrete choice models, would it make sense to substitute a proxy variable for the true state variable? Because of the nonlinearity of the forward-looking discrete choice structure, a naive substitution of the proxy generally biases the estimates of structural parameters, even if the proxy has only an independent error. In this paper, we develop closed-form identification of dynamic discrete choice models when a proxy for an unobserved continuous state variable is available.

Suppose that agent j at time t makes exit decisions $d_{j,t}$ based on its technology $x_{j,t}^*$. Suppose also that we obtain a proxy $x_{j,t} = x_{j,t}^* + \varepsilon_{j,t}$ for the unobserved technology $x_{j,t}^*$ with a classical error $\varepsilon_{j,t}$. If $x_{j,t}^*$ were observable, then identification of the structural parameters of forward-looking agents follows from identification of two auxiliary objects: (1) the conditional choice probability (CCP) denoted by $\Pr(d_t \mid x_t^*)$; and (2) the law of state transition denoted by $f(x_t^* \mid d_{t-1}, x_{t-1}^*)$ (Hotz and Miller, 1993). We show that these two auxiliary objects, $\Pr(d_t \mid x_t^*)$ and $f(x_t^* \mid d_{t-1}, x_{t-1}^*)$, are identified using the proxies $x_{j,t}$ without observing the true states $x_{j,t}^*$.

Indeed, dynamic discrete choice models with unobservables are extensively studied in the literature (e.g., Aguirregabiria and Mira, 2007; Kasahara and Shimotsu, 2009; Arcidiacono and Miller, 2011; Hu and Shum, 2012 – see also the survey by Aguirregabiria and Mira, 2013), but no preceding work handles continuous unobservables like technologies. Our methods allow for continuously distributed unobservables at the expense of the requirement of proxy variables for the unobservables. The use of proxy variables in dynamic structural models is related to Cunha, Heckman, and Schennach (2010) and Todd and Wolpin (2012). Since we estimate the parameters of forward-looking structural models, however, we follow a distinct approach outlined as follows.

In the first step, we identify the CCP and the law of state transition using a proxy variable. For this step, we use an approach related to the closed-estimator of Schennach (2004) and Hu and Sasaki (2015) for nonparametric regression models with measurement errors (cf. Li, 2002), as well as the deconvolution methods (Li and Vuong, 1998; Bonhomme and Robin, 2010). In the second step, the CCP-based method (Hotz, Miller, Sanders and Smith, 1994) is applied to the preliminary non-/semi-parametric estimates of the Markov components to obtain structural parameters of a current-time payoff in a simple closed-form expression. Because of its closed form, our estimator is practical and is free from common implementation problems of convergence and numerical global optimization.

6.2.2 An Overview of the Methodology

In this section, we present a practical guideline of our methodology in the context of the problem of firms' exit decisions based on unobserved technologies. Formal identification and estimation results follow in Sections 6.2.3 and 6.2.4.

Let $d_{j,t} = 1$ indicate the decision of a firm to stay in the market, and let $d_{j,t} = 0$ indicate the decision to exit. The firm chooses $d_{j,t}$ given its technological level $x^*_{j,t}$, and based on its knowledge of the law of stochastic motion of $x^*_{j,t}$. Suppose that the technological state $x^*_{j,t}$ of a firm evolves according to the first-order process

$$x^*_{j,t} = \alpha_t + \gamma_t x^*_{j,t-1} + \eta_{j,t}. \tag{6.45}$$

A firm with its technological level $x^*_{j,t}$ is assumed to receive the current payoff of the affine form $\theta_0 + \theta_1 x^*_{j,t} + \omega^d_{j,t}$ if it is in the market, where $\omega^d_{j,t}$ is the choice-specific private shock.[28] On the other hand, the firm receives zero payoff if it is not in the market. Upon exit from the market, the firm may receive a one-time exit value θ_2, but they will not come back once exited. With this setting, the choice-specific value of the technological state $x^*_{j,t}$ can be written as

$$\begin{aligned} \text{With stay } (d_{j,t}=1): \quad & v_1(x^*_{j,t}) \;=\; \theta_0 + \theta_1 x^*_{j,t} + \omega^1_{j,t} + E\left[\rho V(x^*_{j,t+1};\theta) \mid x^*_{j,t}\right] \\ \text{With exit } (d_{j,t}=0): \quad & v_0(x^*_{j,t}) \;=\; \theta_0 + \theta_1 x^*_{j,t} + \theta_2 + \omega^0_{j,t} \end{aligned}$$

where $\rho \in (0,1)$ is the rate of time preference, $V(\,\cdot\,;\theta)$ is the value function, and the conditional expectation $E[\,\cdot \mid x^*_{j,t}]$ is computed based on the the knowledge of the law (6.45) *including* the distribution of $\eta_{j,t}$.

The fist step toward estimation of the structural parameters is to find a proxy variable $x_{j,t}$ for the unobserved technology $x^*_{j,t}$ with a classical error $\varepsilon_{j,t}$, i.e., $x_{j,t} = x^*_{j,t} + \varepsilon_{j,t}$.

The second step is to estimate the parameters (α_t, γ_t) of the dynamic process (6.45) by the method-of-moment approach, e.g.,

$$\begin{bmatrix} \hat\alpha_t \\ \hat\gamma_t \end{bmatrix} = \begin{bmatrix} 1 & \dfrac{\sum_{j=1}^N x_{j,t-1}\mathbb{1}\{d_{j,t-1}=1\}}{\sum_{j=1}^N \mathbb{1}\{d_{j,t-1}=1\}} \\ \dfrac{\sum_{j=1}^N w_{j,t-1}\mathbb{1}\{d_{j,t-1}=1\}}{\sum_{j=1}^N \mathbb{1}\{d_{j,t-1}=1\}} & \dfrac{\sum_{j=1}^N x_{j,t-1}w_{j,t-1}\mathbb{1}\{d_{j,t-1}=1\}}{\sum_{j=1}^N \mathbb{1}\{d_{j,t-1}=1\}} \end{bmatrix}^{-1} \begin{bmatrix} \dfrac{\sum_{j=1}^N x_{j,t}\mathbb{1}\{d_{j,t-1}=1\}}{\sum_{j=1}^N \mathbb{1}\{d_{j,t-1}=1\}} \\ \dfrac{\sum_{j=1}^N x_{j,t}w_{j,t-1}\mathbb{1}\{d_{j,t-1}=1\}}{\sum_{j=1}^N \mathbb{1}\{d_{j,t-1}=1\}} \end{bmatrix}$$

where $w_{j,t-1}$ is some observed variable that is correlated with $x^*_{j,t-1}$, but uncorrelated with the current technological shock $\eta_{j,t}$ and the idiosyncratic shocks $(\varepsilon_{j,t}, \varepsilon_{j,t-1})$. Examples include lags of the proxy, $x_{j,t-2}$. Note that the proxy $x_{j,t}$ as well as $w_{j,t}$ and the choice $d_{j,t}$ are observed, provided that the firm stays in the market. Because of the interaction with the indicator $\mathbb{1}\{d_{j,t-1} = 1\}$, all the sample moments in the above display are computable from observed data.

[28] For continuous state variables, a more generic structure would be non-parametric, but we consider this parametric form in order to focus on the difficulty related to the unobservability of the state variable. Extending the model to a non-parametric one would entail the integral equation of second kind, the identification of which is developed by Srisuma and Linton (2012).

Having obtained $(\hat{\alpha}_t, \hat{\gamma}_t)$, the third step is to identify the distribution of the idiosyncratic shocks $\varepsilon_{j,t}$. Applying the deconvolution method presented by the references listed in the introduction, we can estimate its characteristic function by the formula

$$\hat{\phi}_{\varepsilon_t}(s) \;=\; \frac{\frac{\sum_{j=1}^N e^{isx_{j,t}} \cdot \mathbb{1}\{d_{j,t}=1\}}{\sum_{j=1}^N \mathbb{1}\{d_{j,t}=1\}}}{\exp\left[\int_0^s \frac{i \cdot \sum_{j=1}^N (x_{j,t+1} - \hat{\alpha}_t) \cdot e^{is' x_{j,t}} \cdot \mathbb{1}\{d_{j,t}=1\}}{\hat{\gamma}_t \cdot \sum_{j=1}^N e^{is' x_{j,t}} \cdot \mathbb{1}\{d_{j,t}=1\}} ds'\right]}.$$

All the moments in this formula involve only the observed variables $x_{j,t}$, $x_{j,t+1}$ and $d_{j,t}$, as opposed to the unobserved true state $x^*_{j,t}$. Thus, they are computable from observed data. Note also that $\hat{\alpha}_t$ and $\hat{\gamma}_t$ are already obtained in the previous step. Hence the right-hand side of this formula is directly computable.

The fourth step is to estimate the CCP, $\Pr(d_t \mid x^*_t)$, of stay given the current technological state x^*_t. Using the estimated characteristic function $\hat{\phi}_{\varepsilon_t}$ produced in the previous step and then applying Schennach (2004) or Hu and Sasaki (2015), we can estimate the CCP by the formula

$$p_t(\xi) \;:=\; \hat{\Pr}(d_{j,t} = 1 \mid x^*_{j,t} = \xi) \;=\; \frac{\int \left(\sum_{j=1}^N \mathbb{1}\{d_{j,t}=1\} \cdot e^{is(x_{j,t}-\xi)}\right) \cdot \hat{\phi}_{\varepsilon_{j,t}}(s)^{-1} \cdot \phi_K(sh) ds}{\int \left(\sum_{j=1}^N e^{is(x_{j,t}-\xi)}\right) \cdot \hat{\phi}_{\varepsilon_{j,t}}(s)^{-1} \cdot \phi_K(sh) ds} \tag{6.46}$$

where ϕ_K is the Fourier transform of a kernel function K and h is a bandwidth parameter. A similar remark to the previous ones applies here: since $d_{j,t}$ and $x_{j,t}$ are observed, this CCP estimate is directly computable using observed data, even though the true state $x^*_{j,t}$ is unobserved.

The fifth step is to estimate the state transition law, $f(x^*_{j,t} \mid x^*_{j,t-1})$. Using the previously estimated characteristic function $\hat{\phi}_{\varepsilon_t}$, we can estimate the state transition law by the formula

$$\hat{f}(x^*_{j,t} = \xi_t \mid x^*_{j,t-1} = \xi_{t-1}) \;=\; \frac{1}{2\pi} \int \frac{\hat{\phi}_{\varepsilon_{j,t-1}}(s\gamma_t) \sum_{j=1}^N e^{is(x_{j,t}-\xi_t)} \cdot e^{is(\alpha_t + \gamma_t \xi_{t-1})}}{\hat{\phi}_{\varepsilon_{j,t}}(s) \sum_{j=1}^N e^{is(\alpha_t + \gamma_t x^*_{j,t-1})}} \cdot \phi_K(sh) ds. \tag{6.47}$$

Finally, by applying our estimated CCP (6.46) and our estimated state transition law (6.47) to the CCP-based method of Hotz and Miller (1993), we can now estimate the structural parameters $\theta = (\theta_0, \theta_1, \theta_2)$. Specifically, if we follow the standard assumption that the choice-specific private shocks independently follow the standard Gumbel (Type I Extreme Value) distribution, then we obtain the restriction

$$\ln p_t(x^*_t) - \ln\left(1 - p_t(x^*_t)\right) \;=\; E[\rho V(x^*_{t+1}; \theta) \mid x^*_t] - \theta_2,$$

where the discounted future value can be written in terms of the parameters θ as

$$E[\rho V(x^*_{t+1};\theta) \mid x^*_t] = E\left[\sum_{s=t+1}^{\infty} \rho^{s-t}\left(\theta_0 + \theta_1 x^*_s + \theta_2(1-p_s(x^*_s)) + \bar{\omega}\right.\right.$$
$$\left.\left. -(1-p_s(x^*_s))\log(1-p_s(x^*_s)) - p_s(x^*_s)\log p_s(x^*_s)\right)\left(\prod_{s'=t+1}^{s-1} p_{s'}(x^*_{s'})\right)\,\middle|\, x^*_t\right],$$

where $\bar{\omega}$ denotes the Euler constant ≈ 0.5772. This conditional expectation can be computed by the state transition law estimated with (6.47), and the CCP $p_t(x^*_t)$ is estimated with (6.46). Hence, with our auxiliary estimates, (6.46) and (6.47), the estimator $\hat{\theta}$ solves the equation

$$\ln \hat{p}_t(x^*_t) - \ln\left(1-\hat{p}_t(x^*_t)\right) = \widehat{E}\left[\sum_{s=t+1}^{\infty} \rho^{s-t}\left(\hat{\theta}_0 + \hat{\theta}_1 x^*_s + \hat{\theta}_2(1-\hat{p}_s(x^*_s)) + \bar{\omega}\right.\right. \tag{6.48}$$
$$\left.\left. -(1-\hat{p}_s(x^*_s))\log(1-\hat{p}_s(x^*_s)) - \hat{p}_s(x^*_s)\log \hat{p}_s(x^*_s)\right)\left(\prod_{s'=t+1}^{s-1} \hat{p}_{s'}(x^*_{s'})\right)\,\middle|\, x^*_t\right] - \hat{\theta}_2 \qquad \text{for all } x^*_t,$$

which can be solved for $\hat{\theta}$ in an OLS-like closed form (cf. Motz, Miller, Sanders and Smith, 1994). The practical advantage of the above estimation procedure is that every single formula is provided with an explicit closed-form expression, and hence does not suffer from the common implementation problems of convergence and global optimization.

Given the structural parameters $\theta = (\theta_0, \theta_1, \theta_2)$ estimated through the above procedure, one can conduct counter-factual policy predictions in the usual manner. For example, consider the policy scenario where the exit value of the current period is reduced by rate r at time t, i.e., the exit value becomes $(1-r)\theta_2$. To predict the number of exits under this experimental setting, we can estimate the counter-factual CCP of stay by the formula

$$\hat{p}^c_t(x^*_t; r) = \frac{\exp\left(\ln \hat{p}_t(x^*_t) - \ln(1-\hat{p}_t(x^*_t)) + r\hat{\theta}_2\right)}{1+\exp\left(\ln \hat{p}_t(x^*_t) - \ln(1-\hat{p}_t(x^*_t)) + r\hat{\theta}_2\right)}.$$

Integrating $\hat{p}^c_t(\,\cdot\,; r)$ over the the unobserved distribution of $x^*_{j,t}$ yields the overall fraction of staying firms, where this unobserved distribution can be in turn estimated by the formula

$$\hat{f}(x^*_{j,t} = \xi_t) = \frac{1}{2\pi}\int \frac{\sum_{j=1}^{N} e^{is(x_{j,t} - x i_t)}}{N \cdot \hat{\phi}_{\varepsilon_{j,t}(s)}} \cdot \phi_K(sh) ds.$$

In this section, we proposed a practical step-by-step guideline of our proposed method. For ease of exposition, this informal overview of our methodology in the current section focused on a specific economic problem and skipped formal assumptions and formal justifications. Readers who are interested in more details of how we derive this methodology may want to go through Sections 6.2.3 and 6.2.4, where we provide formal identification and estimation results in a more general class of forward-looking structural models.

6.2.3 Markov Components: Identification and Estimation

Our basic notations are fixed as follows. A discrete control variable, taking values in $\left\{0, 1, \cdots, \bar{d}\right\}$, is denoted by d_t. For example, it indicates the discrete amounts of lumpy R&D investment, and can take the value of zero which is often observed in empirical panel data for firms. Another example is the binary choice of exit by firms that take into account the future fate of technological progress. An observed state variable is denoted by w_t. It is for example the stock of capital. An unobserved state variable is denoted by x_t^*. It is for example the stock of skills or technologies. Finally, a proxy for x_t^* is denoted by x_t. Throughout this paper, we consider the dynamics of this list of random variables.

Closed-Form Identification of the Markov Components

Our identification strategy is based on the assumptions listed below.

Assumption 6.2.1 (First-Order Markov Process) *The quadruple $\{d_t, w_t, x_t^*, x_t\}$ jointly follows a first-order Markov process.*

This Markovian structure is decomposed into four independent modules as follows.

Assumption 6.2.2 (Independence) *The Markov kernel can be decomposed as follows.*

$$\begin{aligned} & f\left(d_t, w_t, x_t^*, x_t | d_{t-1}, w_{t-1}, x_{t-1}^*, x_{t-1}\right) \\ = \; & f\left(d_t | w_t, x_t^*\right) f\left(w_t | d_{t-1}, w_{t-1}, x_{t-1}^*\right) f\left(x_t^* | d_{t-1}, w_{t-1}, x_{t-1}^*\right) f\left(x_t | x_t^*\right) \end{aligned}$$

where the four components represent

$$\begin{array}{rl} f\left(d_t | w_t, x_t^*\right) & \textit{conditional choice probability (CCP)}; \\ f\left(w_t | d_{t-1}, w_{t-1}, x_{t-1}^*\right) & \textit{transition rule for the observed state variable}; \\ f\left(x_t^* | d_{t-1}, w_{t-1}, x_{t-1}^*\right) & \textit{transition rule for the unobserved state variable}; \textit{and} \\ f\left(x_t | x_t^*\right) & \textit{proxy model.} \end{array}$$

The CCP is the firm's investment or exit decision rule based on the observed capital stocks w_t and the unobserved productivity x_t^* for example. The two transition rules specify how the capital stock w_t and the technology x_t^* co-evolve endogenously with firm's forward-looking decision d_t. The proxy model is a stochastic relation between the true productivity x_t^* and a proxy x_t. Because the state variable x_t^* of interest is unit-less and unobserved, we require a restriction of location- and scale-scale normalization. To this goal, the transition rule for the unobserved state variable and the state-proxy relation are semi-parametrically specified as follows.

Assumption 6.2.3 (Semi-Parametric Restrictions on the Unobservables) *The transition rule for the unobserved state variable and the state-proxy relation are semi-parametrically specified by*

$$f\left(x_t^* | d_{t-1}, w_{t-1}, x_{t-1}^*\right): \qquad x_t^* = \alpha^d + \beta^d w_{t-1} + \gamma^d x_{t-1}^* + \eta_t^d \qquad \textit{if } d_{t-1} = d \tag{6.49}$$

$$f\left(x_t | x_t^*\right): \qquad x_t = x_t^* + \varepsilon_t \tag{6.50}$$

where ε_t and η_t^d have mean zero for each d, and satisfy

$$\begin{array}{lll} \varepsilon_t & \perp & (\{d_\tau\}_\tau, \{x_\tau^*\}_\tau, \{w_\tau\}_\tau, \{\varepsilon_\tau\}_{\tau \neq t}) \qquad \textit{for all } t \\ \eta_t^d & \perp & (d_\tau, x_\tau^*, w_\tau) \qquad\qquad\qquad\qquad\quad \textit{for all } \tau < t \textit{ for all } t. \end{array}$$

Remark 1 *The decomposition in Assumption 6.2.2 and the functional form for the evolution of x_t^* in addition imply that $\eta_t^d \perp w_t$ for all d and t, which is also used to derive our result.*

In case where we consider the discrete choice d_t of investment decisions for example, it is important that the coefficients, $(\alpha^d, \beta^d, \gamma^d)$, are allowed to depend on the amount d of investments since how much a firm invests will likely affect the dynamics of technological evolution. As such, we allow these parameters to have the d superscripts in (6.49). The semiparametric model (6.50) of the state-proxy relation specifies the proxy x_t as a measurement of the latent technology x_t^* with a classical error. Since it is often restrictive in applications, we also discuss how to relax this classical-error assumption in the supplementary note.

By Assumption 6.2.3, closed-form identification of the transition rule for x_t^* and the proxy model for x_t^* follows from identification of the parameters $(\alpha^d, \beta^d, \gamma^d)$ for each d and from identification of the nonparametric distributions of the unobservables, ε_t, x_t^* , and η_t^d for each d. We show that identification of the parameters $(\alpha^d, \beta^d, \gamma^d)$ follows from the empirically testable rank condition stated as Assumption 6.2.4 below.[29] We also obtain identification of the nonparametric distributions of the unobservables, ε_t, x_t^*, and η_t^d, by deconvolution methods under the regularity condition stated as Assumption 6.2.5 below.

Assumption 6.2.4 (Testable Rank Condition) $\Pr(d_{t-1} = d) > 0$ *and the following matrix is nonsingular for each d.*

$$\begin{bmatrix} 1 & E[w_{t-1} \mid d_{t-1} = d] & E[x_{t-1} \mid d_{t-1} = d] \\ E[w_{t-1} \mid d_{t-1} = d] & E[w_{t-1}^2 \mid d_{t-1} = d] & E[x_{t-1} w_{t-1} \mid d_{t-1} = d] \\ E[w_t \mid d_{t-1} = d] & E[w_{t-1} w_t \mid d_{t-1} = d] & E[x_{t-1} w_t \mid d_{t-1} = d] \end{bmatrix}$$

Assumption 6.2.5 (Regularity) *The random variables w_t and x_t^* have bounded conditional first moments given d_t. The conditional characteristic functions of w_t and x_t^* given $d_t = d$ do not vanish on the real line, and is absolutely integrable. The conditional characteristic function of (x_{t-1}^*, w_t) given (d_{t-1}, w_{t-1}) and the conditional characteristic function of x_t^* given w_t are absolutely integrable. Random variables ε_t and η_t^d have bounded first moments and absolutely integrable characteristic functions that do not vanish on the real line.*

The validity of Assumptions 6.2.1, 6.2.2, and 6.2.3 can be discussed with specific economic structures. Assumption 6.2.4 is empirically testable as is the common rank condition in generic econometric contexts. Assumption 6.2.5 consists of technical regularity conditions, but are automatically satisfied by common distribution families, such as the normal

[29] This matrix consists of moments estimable at the parametric rate of convergence, and hence the standard rank tests (e.g., Cragg and Donald, 1997; Robin and Smith, 2000; Kleibergen and Paap, 2006) can be used.

distributions among others. Under this list of five assumptions, we obtain the following closed-form identification result for the four components of the Markov kernel.

Theorem 6.2.1 (Closed-Form Identification) *If Assumptions 6.2.1, 6.2.2, 6.2.3, 6.2.4, and 6.2.5 are satisfied, then the four components $f\left(d_t|w_t,x_t^*\right)$, $f\left(w_t|d_{t-1},w_{t-1},x_{t-1}^*\right)$, $f\left(x_t^*|d_{t-1},w_{t-1},x_{t-1}^*\right)$, $f\left(x_t|x_t^*\right)$ of the Markov kernel $f\left(d_t,w_t,x_t^*,x_t|d_{t-1},w_{t-1},x_{t-1}^*,x_{t-1}\right)$ are identified with closed-form formulas.*

We also show the results with short-hand notations below for convenience of readers. Let $i := \sqrt{-1}$ denote the unit imaginary number. We introduce the Fourier transform operators $\mathcal{F}$ and $\mathcal{F}_2$ defined by

$$\begin{aligned}\mathcal{F}\phi(\xi) &= \frac{1}{2\pi}\int e^{-is\xi}\phi(s)ds \qquad \text{for all } \phi\in L^1(\mathbb{R}) \text{ and } \xi\in\mathbb{R}\\ \mathcal{F}_2\phi(\xi_1,\xi_2) &= \frac{1}{4\pi^2}\int e^{-is_1\xi_1-is_2\xi_2}\phi(s_1,s_2)ds_1ds_2 \quad \text{for all } \phi\in L^1(\mathbb{R}^2) \text{ and } (\xi_1,\xi_2)\in\mathbb{R}^2.\end{aligned}$$

First, with these notations, the CCP (e.g., the conditional probability of choosing the amount d of investment given the capital stock w_t and the technological state x_t^*) is identified in closed form by

$$\Pr\left(d_t=d|w_t,x_t^*\right)=\frac{\mathcal{F}\phi_{(d)x_t^*|w_t}(x_t^*)}{\mathcal{F}\phi_{x_t^*|w_t}(x_t^*)}$$

for each choice $d\in\{0,1,\cdots,\bar{d}\}$, where $\phi_{(d)x_t^*|w_t}(s)$ and $\phi_{x_t^*|w_t}(s)$ are identified in closed form by

$$\phi_{(d)x_t^*|w_t}(s)=\frac{E[\mathbb{1}\{d_t=d\}\cdot e^{isx_t}\mid w_t]}{\phi_{\varepsilon_t}(s)} \qquad \text{and} \qquad \phi_{x_t^*|w_t}(s) = \frac{E[e^{isx_t}\mid w_t]}{\phi_{\varepsilon_t}(s)},$$

respectively, where $\phi_{\varepsilon_t}(s)$ is identified in closed form by

$$\phi_{\varepsilon_t}(s)=\frac{E[e^{isx_t}\mid d_t=d']}{\exp\left[\int_0^s \frac{E[i(x_{t+1}-\alpha^{d'}-\beta^{d'}w_t)\cdot e^{is'x_t}|d_t=d']}{\gamma^{d'}E[e^{is'x_t}|d_t=d']}ds'\right]} \tag{6.51}$$

with any choice d'. For this closed form identifying formula, the parameter vector $(\alpha^d,\beta^d,\gamma^d)^T$ is in turn explicitly identified for each d by the matrix composition

$$\begin{bmatrix} 1 & E[w_{t-1}\mid d_{t-1}=d] & E[x_{t-1}\mid d_{t-1}=d] \\ E[w_{t-1}\mid d_{t-1}=d] & E[w_{t-1}^2\mid d_{t-1}=d] & E[x_{t-1}w_{t-1}\mid d_{t-1}=d] \\ E[w_t\mid d_{t-1}=d] & E[w_{t-1}w_t\mid d_{t-1}=d] & E[x_{t-1}w_t\mid d_{t-1}=d] \end{bmatrix}^{-1} \begin{bmatrix} E[x_t\mid d_{t-1}=d] \\ E[x_tw_{t-1}\mid d_{t-1}=d] \\ E[x_tw_t\mid d_{t-1}=d] \end{bmatrix}.$$

Second, the transition rule for the observed state variable w_t (e.g., the law of motion of capital) is identified in closed form by

$$f\left(w_t|d_{t-1},w_{t-1},x_{t-1}^*\right)=\frac{\mathcal{F}_2\phi_{x_{t-1}^*,w_t|d_{t-1},w_{t-1}}(x_{t-1}^*,w_t)}{\int\mathcal{F}_2\phi_{x_{t-1}^*,w_t|d_{t-1},w_{t-1}}(x_{t-1}^*,w_t)dw_t},$$

where $\phi_{x^*_{t-1},w_t|d_{t-1},w_{t-1}}$ is identified in closed form by

$$\phi_{x^*_{t-1},w_t|d_{t-1},w_{t-1}}(s_1,s_2) \;=\; \frac{E[e^{is_1x_{t-1}+is_2w_t} \mid d_{t-1},w_{t-1}]}{\phi_{\varepsilon_{t-1}}(s_1)}.$$

Third, the transition rule for the unobserved state variable x^*_t (e.g., the evolution of technology) is identified in closed form by

$$f(x^*_t \mid d_{t-1},w_{t-1},x^*_{t-1}) = \mathcal{F}\phi_{\eta^d_t}(x^*_t - \alpha^d - \beta^d w_{t-1} - \gamma^d x^*_{t-1}),$$

where $d := d_{t-1}$ for short-hand notation, and $\phi_{\eta^d_t}$ is identified in closed form by

$$\phi_{\eta^d_t}(s) \;=\; \frac{E[e^{isx_t} \mid d_{t-1}=d]\cdot\phi_{\varepsilon_{t-1}}(s\gamma^d)}{E[e^{is(\alpha^d+\beta^d w_{t-1}+\gamma^d x_{t-1})} \mid d_{t-1}=d]\cdot\phi_{\varepsilon_t}(s)}.$$

Lastly, the proxy model for x^*_t (e.g., the distribution of the idiosyncratic shock as the proxy error) is identified in closed form by

$$f(x_t \mid x^*_t) \;=\; \mathcal{F}\phi_{\varepsilon_t}(x_t - x^*_t),$$

where $\phi_{\varepsilon_t}(s)$ is identified in closed form by (6.51).

In summary, we obtained the four components of the Markov kernel identified with closed-form expressions written in terms of observed data even though we do not observe the true state variable x^*_t. These identified components can be in turn plugged in to the structural restrictions to estimate relevant parameters for the model of forward-looking agents. We present how this step works in Section 6.2.4. Before proceeding with structural estimation, we first show that these identified components of the Markov kernel can be easily estimated by their sample counterparts.

Closed-Form Estimation of the Markov Components

Using the sample counterparts of the closed-form identifying formulas presented in Section 6.2.3, we develop straightforward closed-form estimators of the four components of the Markov kernel. Throughout this section, we assume homogeneous dynamics, i.e., time-invariant Markov kernel, for simplicity. This assumption is not crucial, and can be easily removed with minor modifications. Let h_w and h_x denote bandwidth parameters and let ϕ_K denote the Fourier transform of a kernel function K used for the purpose of regularization.

First, the sample-counterpart closed-form estimator of the CCP $f(d_t \mid w_t, x^*_t)$ is given by

$$\hat{\Pr}\left(d_t = d|w_t,x^*_t\right) \;=\; \frac{\int e^{-isx^*_t}\cdot\hat{\phi}_{(d)x^*_t|w_t}(s)\cdot\phi_K(sh_x)ds}{\int e^{-isx^*_t}\cdot\hat{\phi}_{x^*_t|w_t}(s)\cdot\phi_K(sh_x)ds}$$

for each choice $d \in \{0, 1, \cdots, \bar{d}\}$, where $\hat{\phi}_{(d)x_t^*|w_t}(s)$ and $\hat{\phi}_{x_t^*|w_t}(s)$ are given by

$$\hat{\phi}_{(d)x_t^*|w_t}(s) = \frac{\sum_{j=1}^N \sum_{t=1}^T \mathbb{1}\{D_{j,t} = d\} \cdot e^{isX_{j,t}} \cdot K\left(\frac{W_{j,t}-w_t}{h_w}\right)}{\hat{\phi}_{\varepsilon_t}(s) \cdot \sum_{j=1}^N \sum_{t=1}^T K\left(\frac{W_{j,t}-w_t}{h_w}\right)} \quad \text{and}$$

$$\hat{\phi}_{x_t^*|w_t}(s) = \frac{\sum_{j=1}^N \sum_{t=1}^T e^{isX_{j,t}} \cdot K\left(\frac{W_{j,t}-w_t}{h_w}\right)}{\hat{\phi}_{\varepsilon_t}(s) \cdot \sum_{j=1}^N \sum_{t=1}^T K\left(\frac{W_{j,t}-w_t}{h_w}\right)},$$

respectively, where $\hat{\phi}_{\varepsilon_t}(s)$ is given with any d' by

$$\hat{\phi}_{\varepsilon_t}(s) = \frac{\sum_{j=1}^N \sum_{t=1}^T e^{isX_{j,t}} \cdot \mathbb{1}\{D_{j,t} = d'\} \Big/ \sum_{j=1}^N \sum_{t=1}^T \mathbb{1}\{D_{j,t} = d'\}}{\exp\left[\int_0^s \frac{i \cdot \sum_{j=1}^N \sum_{t=1}^{T-1} (X_{j,t+1} - \alpha^{d'} - \beta^{d'} W_{j,t}) \cdot e^{is' X_{j,t}} \cdot \mathbb{1}\{D_{j,t}=d'\}}{\gamma^{d'} \cdot \sum_{j=1}^N \sum_{t=1}^{T-1} e^{is' X_{j,t}} \cdot \mathbb{1}\{D_{j,t}=d'\}} ds'\right]}. \tag{6.52}$$

While the notations may make things appear sophisticated, all these expressions are straightforward sample-counterparts of the corresponding closed-form identifying formulas provided in the previous section. This CCP estimator is derived in a similar manner to Schennach (2004) and Hu and Sasaki (2015). Large sample properties of this CCP estimator can be found in the supplementary note.

Second, the sample-counterpart closed-form estimator of $f(w_t \mid d_{t-1}, w_{t-1}, x_{t-1}^*)$ is given by

$$\hat{f}\left(w_t | d_{t-1}, w_{t-1}, x_{t-1}^*\right) = \frac{\int\int e^{-s_1 x_{t-1}^* - s_2 w_t} \cdot \hat{\phi}_{x_{t-1}^*, w_t | d_{t-1}, w_{t-1}}(s_1, s_2) \cdot \phi_K(s_1 h_x) \cdot \phi_K(s_2 h_w) ds_1 ds_2}{\int\int\int e^{-s_1 x_{t-1}^* - s_2 w_t} \cdot \hat{\phi}_{x_{t-1}^*, w_t | d_{t-1}, w_{t-1}}(s_1, s_2) \cdot \phi_K(s_1 h_x) \cdot \phi_K(s_2 h_w) ds_1 ds_2 dw_t},$$

where $\hat{\phi}_{x_{t-1}^*, w_t | d_{t-1}, w_{t-1}}$ is given by

$$\hat{\phi}_{x_{t-1}^*, w_t | d_{t-1}, w_{t-1}}(s_1, s_2) = \frac{\sum_{j=1}^N \sum_{t=2}^T e^{is_1 X_{j,t-1} + is_2 W_{j,t}} \cdot \mathbb{1}\{D_{j,t-1} = d_{t-1}\} \cdot K\left(\frac{W_{j,t-1}-w_{t-1}}{h_w}\right)}{\hat{\phi}_{\varepsilon_{t-1}}(s_1) \cdot \sum_{j=1}^N \sum_{t=2}^T \mathbb{1}\{D_{j,t-1} = d_{t-1}\} \cdot K\left(\frac{W_{j,t-1}-w_{t-1}}{h_w}\right)}.$$

Third, the sample-counterpart closed-form estimator of $f(x_t^* \mid d_{t-1}, w_{t-1}, x_{t-1}^*)$ is given by

$$f(x_t^* \mid d_{t-1}, w_{t-1}, x_{t-1}^*) = \frac{1}{2\pi} \int e^{-is(x_t^* - \alpha^d - \beta^d w_{t-1} - \gamma^d x_{t-1}^*)} \cdot \hat{\phi}_{\eta_t^d}(s) \cdot \phi_K(s h_x) ds,$$

where $d := d_{t-1}$ for short-hand notation, and $\hat{\phi}_{\eta_t^d}$ is given by

$$\hat{\phi}_{\eta_t^d}(s) = \frac{\hat{\phi}_{\varepsilon_{t-1}}(s\gamma^d) \cdot \sum_{j=1}^N \sum_{t=2}^T e^{isX_{j,t}} \cdot \mathbb{1}\{D_{j,t-1} = d\}}{\hat{\phi}_{\varepsilon_t}(s) \cdot \sum_{j=1}^N \sum_{t=2}^T e^{is(\alpha^d + \beta^d W_{j,t-1} + \gamma^d X_{j,t-1})} \cdot \mathbb{1}\{D_{j,t-1} = d\}}.$$

Lastly, the sample-counterpart closed-form estimator of $f(x_t \mid x_t^*)$ is given by

$$\hat{f}(x_t \mid x_t^*) \;=\; \frac{1}{2\pi}\int e^{-is(x_t - x_t^*)} \cdot \hat{\phi}_{\varepsilon_t}(s) \cdot \phi_K(sh_x)ds,$$

where $\hat{\phi}_{\varepsilon_t}(s)$ is given by (6.52).

In each of the above four closed-form estimators, the choice-dependent parameters $(\alpha^d, \beta^d, \gamma^d)$ are also explicitly estimated by the matrix composition:

$$\left[\begin{array}{ccc} 1 & \frac{\sum_{j=1}^N \sum_{t=1}^{T-1} W_{jt}\mathbb{1}\{D_{jt}=d\}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \mathbb{1}\{D_{jt}=d\}} & \frac{\sum_{j=1}^N \sum_{t=1}^{T-1} X_{jt}\mathbb{1}\{D_{jt}=d\}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \mathbb{1}\{D_{jt}=d\}} \\ \frac{\sum_{j=1}^N \sum_{t=1}^{T-1} W_{jt}\mathbb{1}\{D_{jt}=d\}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \mathbb{1}\{D_{jt}=d\}} & \frac{\sum_{j=1}^N \sum_{t=1}^{T-1} W_{jt}^2\mathbb{1}\{D_{jt}=d\}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \mathbb{1}\{D_{jt}=d\}} & \frac{\sum_{j=1}^N \sum_{t=1}^{T-1} X_{jt}W_{jt}\mathbb{1}\{D_{jt}=d\}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \mathbb{1}\{D_{jt}=d\}} \\ \frac{\sum_{j=1}^N \sum_{t=1}^{T-1} W_{j,t+1}\mathbb{1}\{D_{jt}=d\}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \mathbb{1}\{D_{jt}=d\}} & \frac{\sum_{j=1}^N \sum_{t=1}^{T-1} W_{jt}W_{j,t+1}\mathbb{1}\{D_{jt}=d\}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \mathbb{1}\{D_{jt}=d\}} & \frac{\sum_{j=1}^N \sum_{t=1}^{T-1} X_{jt}W_{j,t+1}\mathbb{1}\{D_{jt}=d\}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \mathbb{1}\{D_{jt}=d\}} \end{array}\right]^{-1}$$

$$\times \left[\begin{array}{c} \frac{\sum_{j=1}^N \sum_{t=1}^{T-1} X_{j,t+1}\mathbb{1}\{D_{jt}=d\}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \mathbb{1}\{D_{jt}=d\}} \\ \frac{\sum_{j=1}^N \sum_{t=1}^{T-1} X_{j,t+1}W_{jt}\mathbb{1}\{D_{jt}=d\}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \mathbb{1}\{D_{jt}=d\}} \\ \frac{\sum_{j=1}^N \sum_{t=1}^{T-1} X_{j,t+1}W_{j,t+1}\mathbb{1}\{D_{jt}=d\}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \mathbb{1}\{D_{jt}=d\}} \end{array}\right].$$

Each element of the above matrix and vector consists of sample moments of observed data. In fact, not only these matrix elements, but also all the expressions in the estimation formulas provided in this section consist of sample moments of observed data. Thus, despite their apparently sophisticated expressions, computation of these estimators is not that difficult.

6.2.4 Structural Dynamic Discrete Choice Models

In this section, we focus on a class of concrete structural models of forward-looking economic agents. We apply our earlier auxiliary identification results to obtain closed-form estimation of the structural parameters. Agents observe the current state (w_t, x_t^*), where x_t^* is not observed by econometricians. Recall that we deal with a continuous observed state variable w_t and a continuous unobserved state variable x_t^*, and it is not practically attractive to work with nonparametric current-time payoff functions with respect to these continuous state variables. As such, suppose that agents receive the the current payoff of the affine form

$$\theta_d^0 + \theta_d^w w_t + \theta_d^x x_t^* + \omega_{dt}$$

at time t if they make the choice $d_t = d$ under the state (w_t, x_t^*), where ω_{dt} is a private payoff shock at time t that is associated with the choice of $d_t = d$. We may of course extend this affine payoff function to higher-order polynomials at the cost of increased number of parameters. The closed-form identifiability continues to hold as far as the payoff linear with respect to the parameters. Forward-looking agents sequentially make decisions $\{d_t\}$ so as

to maximize the expected discounted sum of payoffs

$$E_t\left[\sum_{s=t}^{\infty}\rho^{s-t}\left(\theta^0_{d_s}+\theta^w_{d_s}w_s+\theta^x_{d_s}x^*_s+\omega_{d_s s}\right)\right],$$

where ρ is the rate of time preference. To conduct counterfactual policy predictions, economists estimate these structural parameters, θ^0_d, θ^w_d, and θ^x_d.

For ease of exposition under many notations, let us focus on the case of binary decision, where d_t takes values in $\{0,1\}$. Since the payoff structure is generally identifiable only up to differences, we normalize one of the intercept parameters to zero, say $\theta^0_1=0$. [30] Furthermore, we assume that ω_{dt} is independently distributed according to the Type I Extreme Value Distribution in order to obtain simple closed-form expressions, although this distributional assumption is not essential. Under this setting, an application of Hotz and Miller's (1993) inversion theorem and some calculations yield the restriction

$$\begin{aligned}\xi(\rho;w_t,x^*_t) \;=\; &\theta^0_0\cdot\xi^0_0(\rho;w_t,x^*_t)+\theta^w_0\cdot\xi^w_0(\rho;w_t,x^*_t)+\theta^w_1\cdot\xi^w_1(\rho;w_t,x^*_t)\\ &+\theta^x_0\cdot\xi^x_0(\rho;w_t,x^*_t)+\theta^x_1\cdot\xi^x_1(\rho;w_t,x^*_t)\end{aligned}\tag{6.53}$$

for all (w_t,x^*_t) for all t, where

$$\begin{aligned}\xi(\rho;w_t,x^*_t) \;=\; &\ln f(1\mid w_t,x^*_t)-\ln f(0\mid w_t,x^*_t)+\\ &\sum_{s=t+1}^{\infty}\rho^{s-t}\cdot E\left[f(0\mid w_s,x^*_s)\cdot\ln f(0\mid w_s,x^*_s)\mid d_t=1,w_t,x^*_t\right]+\\ &\sum_{s=t+1}^{\infty}\rho^{s-t}\cdot E\left[f(1\mid w_s,x^*_s)\cdot\ln f(1\mid w_s,x^*_s)\mid d_t=1,w_t,x^*_t\right]-\\ &\sum_{s=t+1}^{\infty}\rho^{s-t}\cdot E\left[f(0\mid w_s,x^*_s)\cdot\ln f(0\mid w_s,x^*_s)\mid d_t=0,w_t,x^*_t\right]-\\ &\sum_{s=t+1}^{\infty}\rho^{s-t}\cdot E\left[f(1\mid w_s,x^*_s)\cdot\ln f(1\mid w_s,x^*_s)\mid d_t=0,w_t,x^*_t\right]\end{aligned}\tag{6.54}$$

$$\begin{aligned}\xi^0_0(\rho;w_t,x^*_t) \;=\; &\sum_{s=t+1}^{\infty}\rho^{s-t}\cdot E\left[f(0\mid w_s,x^*_s)\mid d_t=1,w_t,x^*_t\right]-\\ &\sum_{s=t+1}^{\infty}\rho^{s-t}\cdot E\left[f(0\mid w_s,x^*_s)\mid d_t=0,w_t,x^*_t\right]-1\end{aligned}\tag{6.55}$$

[30] We may alternatively impose a system of restrictions and augment the least-square estimator following Pesendorfer and Schmidt-Dengler (2007) – see also Sanches, Silva, and Srisuma (2013).

$$\xi_d^w(\rho; w_t, x_t^*) = \sum_{s=t+1}^{\infty} \rho^{s-t} \cdot E\left[f(d \mid w_s, x_s^*) \cdot w_s \mid d_t = 1, w_t, x_t^*\right] - \sum_{s=t+1}^{\infty} \rho^{s-t} \cdot E\left[f(d \mid w_s, x_s^*) \cdot w_s \mid d_t = 0, w_t, x_t^*\right] - (-1)^d \cdot w_t \tag{6.56}$$

$$\xi_d^x(\rho; w_t, x_t^*) = \sum_{s=t+1}^{\infty} \rho^{s-t} \cdot E\left[f(d \mid w_s, x_s^*) \cdot x_s^* \mid d_t = 1, w_t, x_t^*\right] - \sum_{s=t+1}^{\infty} \rho^{s-t} \cdot E\left[f(d \mid w_s, x_s^*) \cdot x_s^* \mid d_t = 0, w_t, x_t^*\right] - (-1)^d \cdot x_t^* \tag{6.57}$$

for each $d \in \{0, 1\}$. See the supplementary note for derivation of (6.53)–(6.57).

In the context of their model, Hotz, Miller, Sanders, and Smith (1994) propose to use (6.53) to construct moment restrictions. We adapt this approach to our model with unobserved state variables. To this end, define the function Q by

$$\begin{aligned} Q(\rho, \theta; w_t, x_t^*) = \xi(\rho; w_t, x_t^*) - \theta_0^0 \cdot \xi_0^0(\rho; w_t, x_t^*) + \theta_0^w \cdot \xi_0^w(\rho; w_t, x_t^*) \\ -\theta_1^w \cdot \xi_1^w(\rho; w_t, x_t^*) - \theta_0^x \cdot \xi_0^x(\rho; w_t, x_t^*) - \theta_1^x \cdot \xi_1^x(\rho; w_t, x_t^*) \end{aligned}$$

where $\theta = (\theta_0^0, \theta_0^w, \theta_1^w, \theta_0^x, \theta_1^x)'$. From (6.53), we obtain the moment restriction

$$E[R(\rho, \theta; w_t, x_t^*)' \; Q(\rho, \theta; w_t, x_t^*)] = 0 \tag{6.58}$$

for any list (row vector) of bounded functions $R(\rho, \theta;\ \cdot,\ \cdot)$. This paves the way for GMM estimation of the structural parameters (ρ, θ). Furthermore, if the rate ρ of time preference is not to be estimated (which is indeed the case in many applications in the literature),[31] then the moment restriction (6.58) can even be written linearly with respect to the structural parameters θ by defining the function R by

$$R(\rho; w_t, x_t^*) = [\xi_0^0(\rho; w_t, x_t^*),\ \xi_0^w(\rho; w_t, x_t^*),\ \xi_1^w(\rho; w_t, x_t^*),\ \xi_0^x(\rho; w_t, x_t^*),\ \xi_1^x(\rho; w_t, x_t^*)].$$

(Note that we can drop the argument θ from this function since none of the right-hand-side components depends on θ.) In this case, the moment restriction (6.58) yields the structural parameters θ by the OLS-like closed-form expression

$$\theta = E\left[R(\rho; w_t, x_t^*)' \; R(\rho; w_t, x_t^*)\right]^{-1} E\left[R(\rho; w_t, x_t^*)' \; \xi(\rho; w_t, x_t^*)\right], \tag{6.59}$$

provided that the following condition is satisfied.

Assumption 6.2.6 (Rank Condition) *$E\left[R(\rho; w_t, x_t^*)' \; R(\rho; w_t, x_t^*)\right]$ is nonsingular.*

While this result is indeed encouraging, an important remark is in order. Since the generated random variables $R(\rho; w_t, x_t^*)$ and $\xi(\rho; w_t, x_t^*)$ depend on the unobserved state

[31] This rate is generally non-identifiable together with the payoffs (Rust, 1994; Magnac and Thesmar, 2002).

variables x_t^* and their unobserved dynamics by their definitional equations (6.54)–(6.57), they need to be constructed properly based on observed variables. This issue can be solved by using the components of the Markov kernel identified with closed-form formulas in Section 6.2.3. Note that the elements of all these generated random variables $R(\rho; w_t, x_t^*)$ and $\xi(\rho; w_t, x_t^*)$ take the form $E[\zeta(w_s, x_s^*) \mid d_t, w_t, x_t^*]$ of the unobserved conditional expectations for various $s > t$, where $\zeta(w_s, x_s^*)$ consists of the explicitly identified CCP $f(d_s \mid w_s, x_s^*)$ and its interactions with w_s, x_s^*, and the log of itself in the formulas (6.54)–(6.57). We can recover these unobserved components in the following manner. If $s = t + 1$, then

$$E[\zeta(w_s, x_s^*) \mid d_t, w_t, x_t^*] = \int\int \zeta(w_{t+1}, x_{t+1}^*) \cdot f(w_{t+1} \mid d_t, w_t, x_t^*) \times f(x_{t+1}^* \mid d_t, w_t, x_t^*)\, dw_{t+1} dx_{t+1}^* \tag{6.60}$$

where $f(w_{t+1} \mid d_t, w_t, x_t^*)$ and $f(x_{t+1}^* \mid d_t, w_t, x_t^*)$ are identified with closed-forms formulas in Theorem 6.2.1. On the other hand, if $s > t + 1$, then

$$\begin{aligned} E[\zeta(w_s, x_s^*) \mid d_t, w_t, x_t^*] = & \sum_{d_{t+1}=0}^{1} \cdots \sum_{d_{s-1}=0}^{1} \int \cdots \int \zeta(w_s, x_s^*) \cdot f(w_s \mid d_{s-1}, w_{s-1}, x_{s-1}^*) \times \\ & f(x_s^* \mid d_{s-1}, w_{s-1}, x_{s-1}^*) \cdot \prod_{\tau=t}^{s-2} f(d_{\tau+1} \mid w_\tau, x_\tau^*) \cdot f(w_{\tau+1} \mid d_\tau, w_\tau, x_\tau^*) \times \\ & \cdot f(x_{\tau+1}^* \mid d_\tau, w_\tau, x_\tau^*)\, dw_{t+1} \cdots dw_s\, dx_{t+1}^* \cdots dx_s^*, \end{aligned} \tag{6.61}$$

where $f(d_t \mid w_t, x_t^*)$, $f(w_{t+1} \mid d_t, w_t, x_t^*)$, and $f(x_{t+1}^* \mid d_t, w_t, x_t^*)$ are identified with closed-form formulas in Theorem 6.2.1.

In light of the explicit decompositions (6.60) and (6.61), the generated random variables $\xi(\rho; w_t, x_t^*)$ and $R(\rho; w_t, x_t^*) = [\xi_0^0(\rho; w_t, x_t^*),\ \xi_0^w(\rho; w_t, x_t^*),\ \xi_1^w(\rho; w_t, x_t^*),\ \xi_0^x(\rho; w_t, x_t^*),\ \xi_1^x(\rho; w_t, x_t^*)]$ defined in (6.54)–(6.57) are identified with closed-form formulas. Therefore, the structural parameters θ are in turn identified in the closed form (6.59). We summarize this result as the following corollary.

Corollary 6.2.1 (Closed-Form Identification of Structural Parameters) *Suppose that Assumptions 6.2.1, 6.2.2, 6.2.3, 6.2.4, 6.2.5, and 6.2.6 are satisfied. Given ρ, the structural parameters θ are identified in the closed form (6.59), where the generated random variables $\xi(\rho; w_t, x_t^*)$ and $R(\rho; w_t, x_t^*) = [\xi_0^0(\rho; w_t, x_t^*),\ \xi_0^w(\rho; w_t, x_t^*),\ \xi_1^w(\rho; w_t, x_t^*),\ \xi_0^x(\rho; w_t, x_t^*),\ \xi_1^x(\rho; w_t, x_t^*)]$ which appear in (6.59) are in turn identified with closed-form formulas through Theorem 6.2.1, (6.54)–(6.57), (6.60), and (6.61).*

Remark 2 *We have left unspecified the measure with respect to which the expectations in (6.58) and thus in (6.59) are taken. The choice is in fact flexible because the original restriction (6.53) holds point-wise for all (w_t, x_t^*). A natural choice is the distribution of (w_t, x_t^*), but it is unobserved. In in the supplementary note, we propose how to evaluate those expectations with respect to this unobserved distribution of (w_t, x_t^*) using observed distribution of (w_t, x_t) while, of course, keeping the closed form formulas. We emphasize that one can pick any distribution with which the testable rank condition of Assumption*

6.2.6 is satisfied.

The closed-form identifying formula for the structural parameters directly translates into a closed-form estimator by substituting the closed-form estimators of the Markov kernel . In the supplementary note, we provide a concrete expression for the closed-form estimator of the structural parameters. Due to the consistency of the Markov component estimators , the consistency of the sample-counterpart estimator of the structural parameters also follows by the continuous mapping theorem. However, asymptotic normality does not hold under mild conditions, as it requires among others sufficiently fast convergence rates of the preliminary Markov component estimators, which do not hold in general.[32]

6.3 Multiple Equilibria in Incomplete Information Games

Xiao (2018) considers a static simultaneous move game, in which player i for $i = 1, 2, \ldots, N$ chooses an action a_i from a choice set $\{0, 1, \ldots, K\}$. Let a_{-i} denote actions of the other players, i.e., $a_{-i} = \{a_1, a_2, \ldots, a_{i-1}, a_{i+1}, \ldots, a_N\}$. The player i's payoff is specified as

$$u_i\left(a_i, a_{-i}, \epsilon_i\right) = \pi_i\left(a_i, a_{-i}\right) + \epsilon_i\left(a_i\right), \tag{6.62}$$

where $\epsilon_i(k)$ for $k \in \{0, 1, \ldots, K\}$ is a choice-specific payoff shock for player i. The object of interest contains the payoff primitives and the equilibrium selection probability. Here we omit other observed state variables. These shocks $\epsilon_i(k)$ are assumed to be private information to player i, while the distribution of $\epsilon_i(k)$ is common knowledge to all the players. A widely used assumption is that the payoff shocks $\epsilon_i(k)$ are independent across all the actions k and all the players i. Let $\Pr\left(a_{-i}\right)$ be player i's belief of other player's actions. The expected payoff of player i from choosing action a_i is then

$$\sum_{a_{-i}} \pi_i\left(a_i, a_{-i}\right) \Pr\left(a_{-i}\right) + \epsilon_i\left(a_i\right) \equiv \Pi_i\left(a_i\right) + \epsilon_i\left(a_i\right) \tag{6.63}$$

The Bayesian Nash Equilibrium is defined as a set of choice probabilities $\Pr\left(a_i\right)$ such that

$$\Pr\left(a_i = k\right) = \Pr\left(\left\{\Pi_i(k) + \epsilon_i(k) > \max_{j \neq k} \Pi_i(j) + \epsilon_i(j)\right\}\right). \tag{6.64}$$

The existence of such an equilibrium is guaranteed by the Brouwer's fixed point theorem. Given an equilibrium, the mapping between the choice probabilities and the expected payoff function has also be established by Hotz and Miller (1993) .

However, multiple equilibria may exist for this game, which means the observed choice probabilities are a mixture from different equilibria. Let e^* denote the index of equilibria.

[32]Specifically, super-smooth distributions cause logarithmic rates of convergence – see Fan (1991), Fan and Truong (1993). Also see the supplementary note for some details. In previous versions of this draft, we used to propose the asymptotic normality under many strong restrictions. In the current draft, we now desist from doing that due to the potential conflicts among the restrictive assumptions that were hard to check.

Under each equilibrium e^*, the players' actions a_i are independent because of the independence assumption of private information, i.e.,

$$a_1 \perp a_2 \perp \ldots \perp a_N | e^*. \tag{6.65}$$

Therefore, the observed correlation among the actions contains information on multiple equilibria. If the support of actions is larger than that of e^*, one can use three players' actions as three measurements for e^*. Otherwise, if there are enough players, one can partition the players into three groups and use the group actions as the three measurements. Comparing with many existing studies on multiple equilibria, using the results for measurement error models makes the nonparametric identification in Xiao (2018) more transparent on why and where the assumptions are imposed and what can and can not be identified.

More detailed description can be found in Ruli Xiao's presentation slides ↗.

6.4 Matching Models with Latent Indices

Diamond and Agarwal (2017) consider an economy containing n workers with characteristics (X_i, ε_i) and n firms described by (Z_j, η_j) for $i, j = 1, 2, \ldots, n$. For example, wages offered by a firm is public information in Z_j or η_j. They assume that the observed characteristics X_i and Z_i are independent of other characteristics ε_i and η_j unobserved to researchers. A firm ranks workers by a human capital index as

$$v(X_i, \varepsilon_i) = h(X_i) + \varepsilon_i. \tag{6.66}$$

The workers' preference for firm j is described by

$$u(Z_j, \eta_j) = g(Z_j) + \eta_j. \tag{6.67}$$

The preferences on both sides are public information in the market. Researchers are interested in the preferences, including functions h, g, and distributions of ε_i and η_j.

A match is a set of pairs that show which firm hires which worker. The observed matches are assumed as outcomes of a pairwise stable equilibrium, where no two agents on opposite sides of the market prefer each other over their matched partners. When the numbers of firms and workers are both large, it can be shown that in the unique pairwise stable equilibrium the firm with the q-th quantile position of preference value, i.e., $F_U(u(Z_j, \eta_j)) = q$ is matched with the worker with the q-th quantile position of the human capital index, i.e., $F_V(v(X_i, \varepsilon_i)) = q$, where F_U and F_V are cumulative distribution functions of u and v.

The joint distribution of (X, Z) from observed pairs then satisfies

$$f(X, Z) = \int_0^1 f(X|q) f(Z|q) \, dq, \tag{6.68}$$

This forms a 2-measurement model. Under the specification of the preferences above, i.e.,

$$\begin{aligned} f(X|q) &= f_{\varepsilon}\left(F_V^{-1}(q) - h(X)\right) \\ f(Z|q) &= f_{\eta}\left(F_U^{-1}(q) - g(Z)\right), \end{aligned} \tag{6.69}$$

the functions h and g can be identified up to a monotone transformation. The intuition is that under suitable conditions if two workers with different characteristics x_1 and x_2 are hired by firms with the same characteristics, i.e., $f_{Z|X}(z|x_1) = f_{Z|X}(z|x_2)$ for all z, then the two workers must have the same observed part of the human capital index, i.e., $h(x_1) = h(x_2)$. A similar argument also holds for function g. In order to further identify the model, Diamond and Agarwal (2017) consider many-to-one matching where one firm may have two or more identical slots for workers. In such a sample, they can observe the joint distribution of (X_1, X_2, Z), where (X_1, X_2) are observed characteristics of the two matched workers. Therefore, they obtain

$$f(X_1, X_2, Z) = \int_0^1 f(X_1|q) f(X_2|q) f(Z|q)\, dq. \tag{6.70}$$

This is a 3-measurement model, for which nonparametric identification is feasible under suitable conditions.

7

Applications in Reduced-Form Econometrics

7.1 Fixed Effects in Panel Data Models

Evdokimov (2010) considers a panel data model as follows: for individual i in period t

$$Y_{it} = g\left(X_{it}, \alpha_i\right) + \xi_{it}, \tag{7.1}$$

where X_{it} is a explanatory variable, Y_{it} is the dependent variable, ξ_{it} is an independent error term, and α_i represents fixed effects. In order to use Kotlarski's identity, he considers the event where $\{X_{i1} = X_{i2} = x\}$ for two periods of data to obtain

$$\begin{aligned} Y_{i1} &= g\left(x, \alpha_i\right) + \xi_{i1}, \\ Y_{i2} &= g\left(x, \alpha_i\right) + \xi_{i2}. \end{aligned} \tag{7.2}$$

Under the assumption that ξ_{it} and α_i are independent conditional on X_{it}, the paper is able to identify the distributions of $g\left(x, \alpha_i\right)$, ξ_{i1} and ξ_{i2} conditional on $\{X_{i1} = X_{i2} = x\}$. That means this identification strategy relies on the static aspect of the panel data model. Assuming that ξ_{i1} is independent of X_{i2} conditional X_{i1}, he then identifies $f\left(\xi_{i1} | X_{i1} = x\right)$, and similarly $f\left(\xi_{i2} | X_{i2} = x\right)$, which leads to identification of the regression function $g\left(x, \alpha_i\right)$ under a normalization assumption.

Shiu and Hu (2013) consider a dynamic panel data model

$$Y_{it} = g\left(X_{it}, Y_{i,t-1}, U_{it}, \xi_{it}\right), \tag{7.3}$$

where U_{it} is a time-varying unobserved heterogeneity or an unobserved covariate, and ξ_{it} is a random shock independent of $(X_{it}, Y_{i,t-1}, U_{it})$. They impose the following Markov-type assumption

$$X_{i,t+1} \perp \left(Y_{it}, Y_{i,t-1}, X_{i,t-1}\right) | \left(X_{it}, U_{it}\right) \tag{7.4}$$

to obtain

$$f_{X_{i,t+1},Y_{it},X_{it},Y_{i,t-1},X_{i,t-1}} = \int f_{X_{i,t+1}|X_{it},U_{it}} f_{Y_{it}|X_{it},Y_{i,t-1},U_{it}} f_{X_{it},Y_{i,t-1},X_{i,t-1},U_{it}} dU_{it}. \quad (7.5)$$

Notice that the dependent variable Y_{it} may represent a discrete choice. With a binary Y_{it} and fixed $(X_{it}, Y_{i,t-1})$, equation (7.5) implies a 2.1-measurement model. Their identification results require users to carefully check conditional independence assumptions in their model because the conditional independence assumption in equation (7.4) is not directly motivated by economic structure.

Freyberger (2018) embeds a factor structure into a panel data model as follows:

$$Y_{it} = g\left(X_{it}, \alpha_i' F_t + \xi_{it}\right), \quad (7.6)$$

where $\alpha_i \in \mathbb{R}^m$ stands for a vector of unobserved individual effects and F_t is a vector of constants. Under the assumption that ξ_{it} for $t = 1, 2, \ldots, T$ are jointly independent conditional on α_i and $X_i = (X_{i1}, X_{i2}, \ldots, X_{iT})$, he obtains

$$Y_{i1} \perp Y_{i2} \perp \ldots \perp Y_{iT} | \left(\alpha_i, X_i\right), \quad (7.7)$$

which forms a 3-measurement model. A useful feature of this model is that the factor structure $\alpha_i' F_t$ provides a more specific identification of the model with a multi-dimensional individual effects α_i than a general argument as in Theorem 2.4.2.

Sasaki (2015) considers a dynamic panel with unobserved heterogeneity α_i and sample attrition as follows:

$$\begin{aligned} Y_{it} &= g\left(Y_{i,t-1}, \alpha_i, \xi_{it}\right) \\ D_{it} &= h\left(Y_{it}, \alpha_i, \eta_{it}\right) \\ Z_i &= \varsigma\left(\alpha_i, \epsilon_i\right) \end{aligned} \quad (7.8)$$

where Z_i is a noisy signal of α_i and $D_{it} \in \{0, 1\}$ is a binary indicator for attrition, i.e., Y_{it} is observed if $D_{it} = 1$. Under suitable restrictions on the error terms, the following conditional independence holds

$$Y_{i3} \perp Z_i \perp Y_{i1} \mid \left(\alpha_i, Y_2 = y_2, D_2 = D_1 = 1\right). \quad (7.9)$$

In the case where α_i is discrete, the model is identified using the results in Theorem 2.4.1. Sasaki (2015) also extends this identification result to more general settings.

Below we provide details in Shiu and Hu (2013).

7.1.1 Background

There are very few papers that provide full nonparametric identification of panel data models in the existing literature. This paper provides sufficient conditions for nonparametric identification of nonlinear dynamic models for panel data with unobserved covariates. These

models take into account the dynamic processes by allowing the lagged value of the dependent variable as one of the explanatory variables as well as containing observed and unobserved permanent (heterogeneous) or transitory (serially-correlated) individual differences. Let Y_{it} be the dependent variable at period t and X_{it} be a vector of observed covariates for individual i. We consider nonlinear dynamic panel data models of the form:

$$Y_{it} = g\left(X_{it}, Y_{it-1}, U_{it}, \xi_{it}\right), \quad \forall i = 1, ..., N; t = 1, ..., T-1, \tag{7.10}$$

where g is an unknown nonstochastic function, U_{it} is an unobserved covariate correlated with other observed explanatory variables (X_{it}, Y_{it-1}), and ξ_{it} stands for a random shock independent of all other explanatory variables $(X_{it}, Y_{it-1}, U_{it})$. The focuses of the above model are on the cases in which the time dimension, T, is fixed and the cross section dimension, N, grows without bound. The unobserved covariate U_{it} may contain two components as follows:

$$U_{it} = V_i + \eta_{it},$$

where V_i is the unobserved heterogeneity or the random effects correlated with the observed covariates X_{it} and η_{it} is an unobserved serially-correlated component.

If the unobserved heterogeneity V_i is treated as a parameter for each i, then both V_i and other unknown parameters need to be estimated for the model (1). When T tends to infinity, the MLE provides a consistent estimator for V_i and other unknown parameters. However, T is fixed and usually small for the panel data model considered here, and therefore, there are not enough observations to estimate these parameters. The model suffers from an incidental parameters problem (Neyman and Scott (1948)). In this paper, the unobserved heterogeneity, V_i, is treated as an unobservable random variable which may be correlated with observed covariates from the same individual. This correlated random effect[1] approach (treating V_i as a random variable correlated with the covariates) allows us to integrate out unobserved variables to construct sieve MLE. This reduces potential computational burden from the incidental parameters problem for sieve MLE estimators in the estimation.[2] The transitory component η_{it} may be a function of all the time-varying RHS variables in the history, i.e., $\eta_{it} = \varphi\left(\{X_{i\tau}, Y_{i\tau-1}, \xi_{i\tau}\}_{\tau=0,1,...,t-1}\right)$ for some function φ.[3] Both observed explanatory variables X_{it} and Y_{it-1} become endogenous if the unobserved covariate U_{it} is ignored. In this paper, we provide assumptions, including high-level injectivity restrictions, under which the distribution of Y_{it} conditional on $(X_{it}, Y_{it-1}, U_{it})$, i.e., $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$, is nonparametrically identified. The nonparametric identification of $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ may lead to that of the general form of our model (7.10) under certain specifications of the distribution

[1] In several studies, random effect means V_i is a random variable independent of the explanatory variables. The discussion here is based on definitions on page 286 of Wooldridge (2010).

[2] The estimation of an individual parameter V_i along with other model parameters leads to an incidental parameters problem. Our sieve MLE has a feature of random effect, treating V_i as a random variable and integrating out a composite unobserved variable to construct a likelihood function. Thus, the proposed sieve MLE has a computational advantage over a fixed effect approach because the individual parameter V_i does not appear in the likelihood function.

[3] By the definition of η_{it}, U_{it} might not only contain the error terms in panels but also some unobserved covariates from the past. Hence, U_{it} denotes an unobserved covariate in this paper.

of the random shock ξ_{it}.

In this paper we adopt the correlated random effect approach for nonlinear dynamic panel data models without specifying the distribution of the initial condition. We treat the unobserved covariate in nonlinear dynamic panel data models as the latent true values in nonlinear measurement error models and the observed covariates as the measurement of the latent true values.[4] We then utilize the identification results in Hu and Schennach (2008), where the measurement error is not assumed to be independent of the latent true values. Their results rely on a unique eigenvalue-eigenfunction decomposition of an integral operator associated with joint densities of observable variables and unobservable variables. Hu and Shum (2012) uses an identification technique described in Carroll et al. (2010). The two identification strategies are different although both use the spectral decomposition of linear operators. The discussion of the difference in the two techniques can be found in Carroll et al. (2010). The conditional independence assumptions in Hu and Shum (2012) are more general than those here but their results require five periods of data in the comparable setting. Our assumptions are more suitable for panel data models. Although some of our assumptions are stronger, our estimator requires only two periods of the dependent variable Y_{it} and three periods of the covariate X_{it}. This advantage is important because semi-nonparametric estimators usually require the sample size to be large.

The strength of our approach is that we provide nonparametric identification of nonlinear dynamic panel data model using two periods of the dependent variable Y_{it} and three periods of the covariate X_{it} without specifying initial conditions. The model may be described by,$f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$, the conditional distribution of the dependent variable of interest for an individual i, Y_{it}, conditional on a lagged value of that variable Y_{it-1}, explanatory variables X_{it}, and an unobserved covariate U_{it}. We show that $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ can be nonparametrically identified from a sample of $\{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}\}$ without parametric assumptions on the distribution of the individuals' dependent variable conditional on the unobserved covariate in the initial period. The main identifying assumption requires that the dynamic process of the covariates X_{it+1} depends on the unobserved covariate U_{it} but is independent of the lagged dependent variables Y_{it}, Y_{it-1}, and X_{it-1} conditional on X_{it} and U_{it}.

The identification of $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ leads to the identification of the general form of our model in equation (7.10). We present below two motivating examples in the existing literature. The specifications in these two types of models can be used to distinguish between dynamic responses to lagged dependent variables, observed covariates, and unobserved covariates. While the state dependence Y_{it-1} reflects that experiencing the event in one period should affect the probability of the event in the next period, the unobserved heterogeneity V_i represents individual's inherent ability to resist the transitory shocks η_{it}.

Example 1 (Dynamic Discrete-choice Model with an Unobserved Covariate): A binary

[4] An ideal candidate for the "measurement" of the latent covariate would be the dependent variable because it is inherently correlated with the latent covariate. However, such a measurement is not informative enough when the dependent variable is discrete and the latent covariate is continuous.

case of dynamic discrete choice models is as follows:

$$Y_{it} = 1\left(X_{it}'\beta + \gamma Y_{it-1} + V_i + \varepsilon_{it} \geq 0\right) \quad \text{with} \quad \forall i = 1, ..., n; t = 1, ..., T-1,$$

where $1(\cdot)$ is the 0-1 indicator function and the error ε_{it} follows an AR(1) process $\varepsilon_{it} = \rho\varepsilon_{it-1} + \xi_{it}$ for some constant ρ. The conditional distribution of the interest is then:

$$f_{Y_{it}|X_{it},Y_{it-1},U_{it}} = \left(1 - F_{\xi_{it}}\left[-\left(X_{it}'\beta + \gamma Y_{it-1} + U_{it}\right)\right]\right)^{Y_{it}} F_{\xi_{it}}\left[-\left(X_{it}'\beta + \gamma Y_{it-1} + U_{it}\right)\right]^{1-Y_{it}},$$

where $F_{\xi_{it}}$ is the CDF of the random shock ξ_{it}, $U_{it} = V_i + \eta_{it}$, and $\eta_{it} = \rho\varepsilon_{it-1}$. Empirical applications of the dynamic discrete-choice model above have been studied in a variety of contexts, such as health status (Contoyannis et al. (2004), Halliday (2002)), brand loyalty (Chintagunta et al. (2001)), welfare participation (Chay et al. (2001)), and labor force participation (Heckman and Willis (1977), Hyslop (1999)). Among these studies, the intertemporal labor participation behavior of married women is a natural illustration of the dynamic discrete choice model. In such a model, the dependent variable Y_{it} denotes the t-th period participation decision and the covariates X_{it} are the non-labor income or other observable characteristics in that period. The heterogeneity V_i is the unobserved individual skill level or motivation, while the idiosyncratic disturbance ξ_{it} denotes unexpected change of child-care cost or fringe benefit for married women from working. Heckman (1978, 1981a,b) has termed the presence of Y_{it-1} "true" state dependence and V_i "spurious" state dependence.

Example 2 (Dynamic Censored Model with an Unobserved Covariate): In many applications, we may have

$$Y_{it} = \max\left\{X_{it}'\beta + \gamma Y_{it-1} + V_i + \varepsilon_{it}, 0\right\} \quad \text{with} \quad \forall i = 1, ..., n; t = 1, ..., T-1,$$

with $\varepsilon_{it} = \rho\varepsilon_{it-1} + \xi_{it}$. It follows that

$$f_{Y_{it}|X_{it},Y_{it-1},U_{it}} = F_{\xi_{it}}\left[-\left(X_{it}'\beta + \gamma Y_{it-1} + U_{it}\right)\right]^{\mathbf{1}(Y_{it}=0)} f_{\xi_{it}}\left[Y_{it} - X_{it}'\beta - \gamma Y_{it-1} - U_{it}\right]^{\mathbf{1}(Y_{it}>0)}. \tag{7.11}$$

where $F_{\xi_{it}}$ and $f_{\xi_{it}}$ are the CDF and the PDF of the random shock ξ_{it} respectively. The dependent variable Y_{it} may stand for the amount of insurance coverage chosen by an individual or a firm's expenditures on R&D. In each case, an economic agent solves an optimization problem and $Y_{it} = 0$ may be an optimal corner solution. For this reason, this type of censored regression models is also called a corner solution model or a censored model with lagged censored dependent variables.[5] Honoré (1993) and Honoré and Hu (2004) use a method of moments framework to estimate the model without making distributional assumptions about V_i.

Based on our nonparametric identification results, we propose a semi-parametric sieve MLE for the model. We show the consistency of our estimator and the asymptotic normality

[5]This setting rules out certain types of data censoring. For example, if the censoring is due to top-coding, then it makes sense to consider a lagged value of the latent variable, i.e., $Y_{it}^* = X_{it}'\beta + \gamma Y_{it-1}^* + v_i + \varepsilon_{it}$ and $Y_{it} = \max[Y_{it}^*, c_t]$. This top-coded dynamic censored model has been considered in Hu (2002).

of its parametric components. The finite sample properties of the proposed sieve MLE are investigated through Monte Carlo simulations of dynamic discrete choice models and dynamic censored models. Our empirical application focuses on how the labor participation decisions of married women respond to their previous participation states, fertility decisions, and non-labor incomes. We develop and test a variety of dynamic econometric models using a seven year longitudinal sample from the Panel Study of Income Dynamics (PSID) in order to compare the results with those in Hyslop (1999). In the empirical application, we examine three different estimation specifications, i.e., a static probit model, a maximum simulated likelihood (MSL) estimator, and the sieve MLE estimator. Our results find a large significant state dependence of labor force participation, smaller significant negative effects on non-labor income variables, and also negative effects of children age 0-2 in the current period and past period.

The paper is organized as follows. Section 2 provides a brief review of studies in the context of dynamic panel data models. We present the nonparametric identification of nonlinear dynamic panel data models in Section 3. Section 4 discusses our proposed sieve MLE. Section 5 provides the Monte Carlo study. Section 6 presents an empirical application describing the intertemporal labor participation of married women. Section 7 concludes. Appendices include proofs of consistency and asymptotic normality of the proposed sieve MLE and discussions on how to impose restrictions on sieve coefficients in the sieve MLE.

7.1.2 Related Studies

In the econometric literature, there are two approaches to tackling the unobserved heterogeneity V_i: random effects and fixed effects. In the fixed effect approach, much attention has been devoted to linear models with an additive unobserved effect. The problem can be solved by first applying an appropriate transformation to eliminate the unobserved effect and then implementing instrument variables (IV) in a generalized method of moments (GMM) framework. Anderson and Hsiao (1982), Arellano and Bond (1991), Arellano and Bover (1995) and Ahn and Schmidt (1995) employ an IV estimator on a transformation equation through first-differencing. Eliminating the unobserved effects is notably more difficult in nonlinear models, and some progress has been made in this area. Chamberlain (1980, 1984) consider a conditional likelihood approach for logit models. Honoré and Kyriazidou (2000) generalize the conditional probability approach to estimate the unknown parameters without formulating the distribution of the unobserved individual effects or the probability distribution of the initial observations for certain types of discrete choice logit models. Their results rely on matching the explanatory variables in different time-periods. Honoré (1993), Hu (2002) and Honoré and Hu (2004) obtain moment conditions for estimating dynamic censored regression panel data models. Altonji and Matzkin (2005) develop two estimators for panel data models with nonseparable unobservable errors and endogenous explanatory variables.

On the other hand, it is often appealing to take a random effect specification by making assumptions on the distribution of the individual effects. The main difficulty of this ap-

proach is the so-called initial conditions problem.[6] With a relatively short panel, the initial conditions have a very strong impact on the entire path of the observations, but they may not be observed in the sample. One remedy to this problem is to specify the distribution of the initial conditions given the unobserved heterogeneity. The drawbacks of this approach are that the corresponding likelihood functions typically involve high order integration and that misspecification of the distributions generally results in inconsistent parameter estimates. The associated computational burden of high order integration has been reduced significantly by recent advances in simulation techniques.[7] Hyslop (1999) analyzes the intertemporal labor force participation behavior of married women using maximum simulated likelihood (MSL) estimator to simulate the likelihood function of dynamic probit models with a nontrivial error structure. Wooldridge (2005) suggests a general method for handling the initial conditions problem by using a joint density conditional on the strictly exogenous variables and the initial condition. Honoré and Tamer (2006) relax the distributional assumption of the initial condition and calculate bounds on parameters of interest in panel dynamic discrete choice models. Evdokimov (2010) considers a nonparametric panel data model with nonadditive unobserved heterogeneity: $Y_{it} = m(X_{it}, V_i) + \varepsilon_{it}$ where individual-specific effects are allowed to be correlated with the covariates in an arbitrary manner. That model has a different focus from ours since our model includes lags of the endogenous dependent variable Y_{it-1} and a nonadditive ε_{it}.

While the proposed model (7.10) focuses on nonlinear dynamic panel data models, there are several studies on panel data models that are close in spirit to our work. Chernozhukov et al. (2009) derive bounds for marginal effects in nonlinear panel models and show that they can tighten rapidly as the number of time series observations grows. They also provide two novel inference methods that produce uniformly valid confidence regions in large samples. Hoderlein and White (2009) consider identification of marginal effects in general nonseparable models with unrestricted correlated unobserved effects and without lagged dependent variables, even if there are only two time periods. Arellano and Bonhomme (2009) provide a characterization of the class of weights for nonlinear panel data models that produce first-order unbiased estimators. Although the focus of the models in this paper is on

[6]The random effect approach for dynamic models requires the specification on the initial conditions of the process. Specifically, consider a special case of our model (7.10), dynamic discrete choice models without observed covariates X_{it}, in the following form:

$$Y_{it} = 1(\gamma Y_{it-1} + V_i + \xi_{it} \geq 0).$$

Then the conditional distribution $f_{Y_{it}|Y_{it-1},V_i}$ can be specified and the corresponding likelihood function has the structure

$$\mathcal{L} = \int f_{Y_{i0}|V_i} \prod_{t=1}^{T-1} f_{Y_{it}|Y_{it-1},V_i} f_{V_i} dv_i,$$

where $f_{Y_{i0}|V_i}$ denotes the marginal probability of Y_{i0} given V_i. If the process is not observed from the start then the initial state for individual i, y_{i0} cannot be assumed fixed. However, it is not clear how to derive the initial condition $f_{Y_{i0}|V_i}$ from $f_{Y_{it}|Y_{it-1},V_i}$ so it could be internally inconsistent across different time periods if the evolution of these two process cannot be connected. Heckman (1981b) suggested that using a flexible functional form to approximate the initial conditions.

[7]See Gourieroux and Monfort (1993), Hajivassiliou (1993), Hajivassiliou and Ruud (1994) and Keane (1993) for the reviews of the literature.

the fixed time dimension, the results can be generalized to large T cases.

In this paper, we provide nonparametric identification of nonlinear dynamic panel data models with unobserved covariates, show the models are identified using only two periods of the dependent variable Y_{it} and three periods of the covariate X_{it} without initial conditions assumptions, and propose a sieve MLE estimator. The advantages of our results include that the point identification results are nonparametric and global, the model is quite general comparing with the existing ones and makes use of the recently developed techniques, and the proposed sieve estimator is known to be convenient in computation. Meanwhile, our results have their disadvantages. The general nonparametric identifiction requires high-level technical assumptions. In particular, the injectivity assumption is not testable and its implication is still an active reseach area. The proposed sieve estimator also has its known shortcomings, such as the difficulty in choosing nuisance parameters.

7.1.3 Nonparametric Identification

Main Assumptions

In this section, we present the assumptions under which the distribution of the dependent variable Y_{it} conditional on Y_{it-1}, covariates X_{it}, and the unobserved covariate U_{it}, i.e., $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$, is nonparametrically identified. As discussed above, some of our assumptions are high-level because we are providing nonparametric identification of the model. We assume

Assumption 7.1.1 *(Exogenous shocks) Assume* $f_{Y_{it}|X_{it},Y_{it-1},X_{it-1},U_{it}} = f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$.

A sufficient condition for Assumption 3.1 is that the random shock ξ_{it} is independent of $\xi_{i\tau}$ for any $\tau \neq t$ and $\{X_{i\tau}, U_{i\tau}\}$ for any $\tau \leq t$. Given Eq. (1), the condition $f_{Y_{it}|X_{it},Y_{it-1},X_{it-1},U_{it}} = f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ holds if the random shock ξ_{it} is independent of the covariate X_{it-1}. This assumption can be called an exogenous shocks condition. As shown in the two examples above, this sufficient assumption has been used in many existing studies.

Both ξ_{it} and U_{it} are scalar unobservables in the latent variable formulation of the dependent variable Y_{it} and account for the particular error structure in the formulation. While ξ_{it} is an exogenous random shock in period t, $U_{it} = V_i + \eta_{it}$ is the sum of the time-invariant heterogeneity and a function of all time-varying variables in the past.

The exogeneity of ξ_{it} can be relaxed to allow some dependence between ξ_{it} and (X_{it}, Y_{it-1}). For example, for some positive function h, write $\xi_{it} = h\left(X_{it}, Y_{it-1}\right)^{1/2} e_{it}$ for an exogenous random shock e_{it} with unit variance. Hence, ξ_{it} contains heteroskedasticity and $\mathrm{Var}(\xi_{it}|X_{it}, Y_{it-1}) = h\left(X_{it}, Y_{it-1}\right)$. In this case, the conditional distribution of the interest in Example 1 changes into:

$$
\begin{aligned}
&f_{Y_{it}|X_{it},Y_{it-1},U_{it}} \\
&= \left(1 - F_{\xi_{it}}\left[\frac{-\left(X_{it}'\beta + \gamma Y_{it-1} + U_{it}\right)}{h\left(X_{it}, Y_{it-1}\right)^{1/2}}\right]\right)^{Y_{it}} F_{\xi_{it}}\left[\frac{-\left(X_{it}'\beta + \gamma Y_{it-1} + U_{it}\right)}{h\left(X_{it}, Y_{it-1}\right)^{1/2}}\right]^{1-Y_{it}}.
\end{aligned}
$$

Making ξ_{it} heteroskedastic generalizes the functional form of the dynamic panel data models considered in this paper. However, for simplicity we assume ξ_{it} is exogenous with a constant variance.

The existence of the exogenous random shock ξ_{it} in the error term of the latent variable formulation means that (X_{it}, Y_{it-1}) fully capture the dynamics conditional on U_{it} since further lags of Y_{it-1} or lags of X_{it} are not important once $(X_{it}, Y_{it-1}, U_{it})$ have been controlled for. To some extent, Assumption 7.1.1 has assumed dynamic completeness since

$$f_{Y_{it}|X_{it},Y_{it-1},U_{it}} = f_{Y_{it}|X_{it},Y_{it-1},U_{it},X_{it-1},Y_{it-2},U_{it-1},...,X_{i1},Y_{i0},U_{i1}}, \quad t = 1, ..., T-1,$$

and once U_{it} is controlled for, no past values of X_{it} or Y_{it-1} appear in the conditional density in the RHS of the above equation.

We simplify the evolution of the observed covariates X_{it} as follows:

Assumption 7.1.2 *(Covariate evolution) Assume the covariate evolution satisfies the equation* $f_{X_{it+1}|Y_{it},X_{it},Y_{it-1},X_{it-1},U_{it}} = f_{X_{it+1}|X_{it},U_{it}}$.

Note that the assumption can be written as $X_{it+1} \perp (Y_{it}, Y_{it-1}, X_{it-1}) | (X_{it}, U_{it})$ and the lagged effects of Y_{it} such as $Y_{it-1}, Y_{it-2}, ...$ enter the evolution of X_{it+1} through the unobserved covariate U_{it}. A sufficient condition for Assumption 7.1.2 is that X_{it+1} is strictly exogenous and follows a first order Markov, conditional on U_{it}. Another sufficient condition for Assumption 3.2 is constituted of three steps, (i)(Markov evolution of X_{it+1}) $f_{X_{it+1}|Y_{it},X_{it},Y_{it-1},X_{it-1},U_{it}} = f_{X_{it+1}|Y_{it},X_{it},U_{it}}$, (ii)(No impact of ξ_{it} on X_{it+1}) $f_{X_{it+1}|Y_{it},Y_{it-1},X_{it},U_{it}} = f_{X_{it+1}|Y_{it-1},X_{it},U_{it}}$, and (iii)(Limited feedback) $f_{X_{it+1}|Y_{it-1},X_{it},U_{it}} = f_{X_{it+1}|X_{it},U_{it}}$.

The first step (i) is a Markov-type assumption $f_{X_{it+1}|Y_{it},X_{it},Y_{it-1},X_{it-1},U_{it}} = f_{X_{it+1}|Y_{it},X_{it},U_{it}}$, which implies that the evolution of the observed covariate X_{it+1} only depends on all the explanatory variables in the previous period (Y_{it}, X_{it}, U_{it}). The implication of the Markov assumption is that while the time-varying component of U_{it}, η_{it}, captures all the serially-correlated variation in the process of X_{it+1}, the corresponding time-invariant component V_i controls the time-invariant part of X_{it+1}. If X_{it+1} contains a time-invariant component other than V_i then the Markov assumption may fail. For example, suppose that we have[8]

$$X_{it+1} = \rho X_{it} + W_i + V_i + v_{it},$$

where v_{it} are i.i.d, and a latent W_i is not perfectly correlated with V_i. In this case, given U_{it}, X_{it-1} will contain some information about W_i, even given X_{it}. Thus, X_{it-1} can be informative on X_{it+1} given (Y_{it}, X_{it}, U_{it}) and the Markov condition does not hold. However, the composite error U_{it} is a scalar unobservable in the latent variable formulation of the dependent variable Y_{it} and should also take account of the variation of X_{it}. If the time-varying component of U_{it} contains v_{it} and its time-invariant component has $W_i + V_i$, the Markov assumption may hold. Our assumption rules out the situation that the evolution of X_{it} depends on other time-invariant element not in the latent variable formulation of Y_{it}.

[8]We thank an anonymous referee for suggesting this example.

The second step (ii) is that conditional on Y_{it-1}, X_{it} and U_{it}, X_{it+1} is independent of the exogenous shock ξ_{it}. Since U_{it} is a function of all past shocks $\{\xi_{i\tau}\}_{\tau<t}$, this step only excludes the immediate effect of the current shock ξ_{it} on the future covariate X_{it+1}.[9] This implies that $f_{X_{it+1}|X_{it},Y_{it-1},U_{it},\xi_{it}} = f_{X_{it+1}|X_{it},Y_{it-1},U_{it}}$. The third step (iii) is a limited feedback assumption, i.e., $f_{X_{it+1}|X_{it},Y_{it-1},U_{it},} = f_{X_{it+1}|X_{it},U_{it}}$ which rules out direct feedback from the lagged dependent variable Y_{it-1} on the future value of the observed covariate X_{it+1}. The effect of Y_{it-1} on X_{it+1} is indirectly through X_{it}, and U_{it}.

Overall, Assumption 7.1.2 implies that conditional on X_{it} and U_{it}, X_{it+1} is independent of the exogenous shock ξ_{it}. In other words, conditional on the past information, the future covariate X_{it+1} rules out the immediate effect of the current shock ξ_{it} of the dependent variable Y_{it}.

Let $\mathcal{L}^p(\mathcal{X}), 1 \leq p < \infty$ stand for the space of function $h(\cdot)$ with $\int_{\mathcal{X}} |h(x)|^p dx < \infty$. Suppose $\mathcal{X}_t$, and $\mathcal{U}_t$ be the supports of the random variables X_{it} and U_{it}, respectively. For any $1 \leq p \leq \infty$ and we define operators as follows: for any given (x_{it}, y_{it-1}),

$$L_{X_{it+1},x_{it},y_{it-1},X_{it-1}} : \mathcal{L}^p(\mathcal{X}_{t-1}) \to \mathcal{L}^p(\mathcal{X}_{t+1})$$
$$(L_{X_{it+1},x_{it},y_{it-1},X_{it-1}}h)(u) = \int f_{X_{it+1},X_{it},Y_{it-1},X_{it-1}}(u, x_{it}, y_{it-1}, x)h(x)dx,$$

and for any given x_{it},

$$L_{X_{it+1}|x_{it},U_{it}} : \mathcal{L}^p(\mathcal{U}_t) \to \mathcal{L}^p(\mathcal{X}_{t+1})$$
$$(L_{X_{it+1}|x_{it},U_{it}}h)(x) = \int f_{X_{it+1}|X_{it},U_{it}}(x|x_{it}, u)h(u)du,$$

Assumption 7.1.3 *(Invertibility) For any* $(x_{it}, y_{it-1}) \in \mathcal{X}_{it} \times \mathcal{Y}_{it-1}$, $L_{X_{it+1},x_{it},y_{it-1},X_{it-1}}$ *and* $L_{X_{it+1}|x_{it},U_{it}}$ *are invertible.*

This is a high-level assumption, which is hard to avoid for nonparametric identification. Intuitively, this assumption guarantees that the observables contain enough information on the unobserved covariate U_{it} and the covariates in period $t+1$, X_{it+1}, depend on X_{it}. However, the invertibility of $L_{X_{it+1},x_{it},y_{it-1},X_{it-1}}$, which is equivalent to a completeness condition on an observed distribution $f_{X_{it+1},X_{it},Y_{it-1},X_{it-1}}$. Its testibility is shown in Canay et al. (2013b) and Freyberger (2017).

If an operator is constructed by a density of independent variables, the operator certainly fails to be invertible. Since $f_{X_{it+1},X_{it},Y_{it-1},X_{it-1}}$ is the density of correlated variables, it provide at least some justification for the completeness property.[10] Thus, the invertibility may require functional form restrictions on $f_{X_{it+1},X_{it},Y_{it-1},X_{it-1}}$. For example, if $\mathcal{X}_{t+1}$ con-

[9]The assumption imposes some restriction to regressors in panel data setting. For example, suppose that $U_{it} = V_i$. The assumption that X_{it+1} is independent of ξ_{it} given X_{it} and V_i implies that $E[X_{it+1}\xi_{it}] = 0$. If the future covariate X_{it+1} is predetermined, in the sense that $E[X_{it+1}\xi_{is}] \neq 0$ for $s < t+1$ and zero otherwise, then the assumption fails when the X_{it+1} is predetermined. However, the assumption permits a weaker version of a predetermined variable such as $E[X_{it+1}\xi_{is}] \neq 0$ for $s < t$ and zero otherwise.

[10]That the variables X_{it+1}, X_{it} ,Y_{it-1}, and X_{it-1} are highly correlated can be justified by the fact that most variables in economics are correlated across time which reveal a pattern of serial correlation or autocorrelation.

tains an open set then $f_{X_{it+1},X_{it},Y_{it-1},X_{it-1}} = \phi(X_{it-1} - \alpha_1 X_{it+1} - \alpha_2 X_{it} - \alpha_3 Y_{it-1})$ satisfies Assumption 7.1.3 where ϕ is the standard normal pdf and $\alpha_i \neq 0$.[11] Besides a linear process, another example may be that $f_{X_{it+1},X_{it},Y_{it-1},X_{it-1}}$ belongs to an exponential family. Given a fixed (x_{it}, y_{it-1}). Suppose that

$$\begin{aligned} &f_{X_{it+1},x_{it},y_{it-1},X_{it-1}} \\ &= s(x_{it}, y_{it-1}, X_{it-1}) t(X_{it+1}, x_{it}, y_{it-1}) \exp\left[\mu(X_{it+1}, x_{it}, y_{it-1}) \tau(x_{it}, y_{it-1}, X_{it-1})\right] \end{aligned}$$

where $s(x_{it}, y_{it-1}, X_{it-1}) > 0$, $\tau(x_{it}, y_{it-1}, X_{it-1})$ is one-to-one in X_{it-1}, and support of $\mu(X_{it+1}, x_{it}, y_{it-1}) \in \mathcal{X}_{t+1}$ is an open set. Theorem 2.2 in Newey and Powell (2003) shows the family of the joint density functions $\{f_{X_{it+1},x_{it},y_{it-1},X_{it-1}} : X_{it+1} \in \mathcal{X}_{t+1}\}$ is complete over $\mathcal{L}^p(\mathcal{X}_{t-1})$ for each (x_{it}, y_{it-1}).[12] This also implies the invertibility of $L_{X_{it+1},x_{it},y_{it-1},X_{it-1}}$ in Assumption 7.1.2.

On the other hand, the invertibility of $L_{X_{it+1}|x_{it},U_{it}}$ requires the covariates in period $t+1$, X_{it+1}, contains enough information on the unobserved covariate U_{it} conditional on X_{it}. Hahn (2001) considers a dynamic logit model with individual effects where the regressors include the lag dependent variable, time dummies and possibly other strictly exogenous variables. He shows that the semi parametric information bound for any estimator of the state dependence coefficient is zero. Our results do not cover the dynamic logit model in Hahn (2001) because the invertibility of $L_{X_{it+1}|x_{it},U_{it}}$ in Assumption 3.3 requires some dependence between U_{it} and X_{it+1}. If X_{it+1} only contains time dummies and possibly other strictly exogenous variables, the condition will fail to hold. This is intuitive: the existence of a degree of dependence between U_{it} and X_{it+1} allows us to control the unobservable U_{it} from the observable X_{it+1}. It reflects the methodology of our identification method that provides an alternative way to deal with an unobservable term inside a nonlinear econometric model, tackling down an unobserved effect with an observable correlated covariate instead of eliminating the unobserved effect by transformations. For example, we may have $X_{it+1} = X_{it} + U_{it} + h(X_{it})\epsilon_{it}$, where ϵ_{it} is independent of X_{it} and U_{it} and has a nonvanishing characteristic function on the real line. We use X_{it+1} instead of Y_{it+1} for the information on U_{it} because the dependent variable Y_{it+1} is discrete and U_{it} is continuous in many interesting applications. In that case, the operator mapping from functions of U_{it} to those of Y_{it+1} cannot be invertible. Additionally, when Y_{it+1} is continuous, it would be more reasonable to impose invertibility on the operator mapping from functions of U_{it} to those of Y_{it+1}, while U_{it} or V_i is allowed to be independent of the observed covariates X_{it}.[13] Neces-

[11] The result is from Theorem 2.3 in Newey and Powell (2003). Suppose that the distribution of x conditional on z is $N(a + bz, \sigma^2)$ for $\sigma^2 > 0$ and the support of z contains an open set, then the integral operator corresponding to $\frac{1}{\sigma}\phi(\frac{x-a-bz}{\sigma})$ is invertible from $\mathcal{L}^p(\mathcal{X})$ to $\mathcal{L}^p(\mathcal{Z})$ where ϕ is the standard normal PDF. There are more detailed discussions and general conditions for an invertible integral operator or complete conditional distributions in $\mathcal{L}^p(\mathcal{X})$ in Hu and Shiu (2018).

[12] The whole statement of the theorem is the following: *Let $f(x|z) = s(x)t(z)\exp[\mu(z)\tau(x)]$, where $s(x) > 0$, $\tau(x)$ is one-to-one in x, and support of $\mu(z)$, $\mathcal{Z}$, is an open set, then $E[h(\cdot)|z] = 0$ for any $z \in \mathcal{Z}$ implies $h(x) = 0$ almost everywhere in $\mathcal{X}$; equivalently, the family of conditional density functions $\{f(x|z) : z \in \mathcal{Z}\}$ is complete in $L^p(\mathcal{X})$.*

[13] Assumption 7.1.3 requires $L_{X_{it+1}|x_{it},U_{it}}$ is invertible and it demands the unobservable U_{it} to be correlated with the observed X_{it+1}. This case is complementary to the existing models where U_{it} is independent of

sary conditions for Assumption 7.1.3 include that $f_{X_{it+1},Y_{it-1},X_{it}|X_{it-1}} \neq f_{X_{it+1},Y_{it-1},X_{it}}$ and $f_{X_{it+1}|X_{it},U_{it}} \neq f_{X_{it+1}|X_{it}}$. These necessary conditions rule out the case where X_{it+1} and X_{it-1} are independent or X_{it+1} and U_{it} are independent. In other words, Assumption 7.1.3 permits the existence of serial correlation among X_{it} and correlation between X_{it+1} and U_{it}.

The invertibility of the integral operator $L_{X_{it+1}|x_{it},U_{it}}$ is equivalent to saying that the family $\{f_{X_{it+1}|X_{it},U_{it}}(x_{it+1}|x_{it},u_{it}) : x_{it+1} \in \mathcal{X}_{t+1}\}$ is complete over $\mathcal{L}^p(\mathcal{U}_t)$. Hu and Shiu (2011) showed that if the conditional density $f(x|z)$ can form a linearly independent sequence and coincides with a known complete density at a limit point in the support of z, then $f(x|z)$ itself is complete. They also provide examples of complete families other than trivial linear/exponential family cases. For example, suppose ϕ is the standard normal pdf, consider

$$f(x|z) = \lambda(z)\, h(x|z) + [1 - \lambda(z)]\, \phi(x - z), \tag{7.12}$$

which is a mixture of two continuous conditional densities, h and ϕ, and the weight λ in the mixture depends on z.[14] Sufficient conditions for the completeness of $f(x|z)$ are (i) $\lim_{z_k \to z_0} \lambda(z) = 0$; and (ii) $\lim_{x \to -\infty} \frac{h(x|z_k)}{\phi(x - z_k))} < \infty$. Following this result, construct

$$\begin{aligned} &f_{X_{it+1}|X_{it},U_{it}}(x_{it+1}|x_{it},u_{it}) \\ &= \lambda(x_{it+1})\, h(x_{it+1},x_{it},u_{it}) + [1 - \lambda(x_{it+1})]\, \phi(x_{it+1} - \psi(x_{it}) - u_{it}), \end{aligned}$$

with $\lim_{x_{it+1,k} \to x_{it+1,0}} \lambda(x_{it+1}) = 0$; and (ii) $\lim_{u_{it} \to -\infty} \frac{h(x_{it+1,k},x_{it},u_{it})}{\phi(x_{it+1,k} - \psi(x_{it}) - u_{it})} < \infty$. The completeness of $\{f_{X_{it+1}|X_{it},U_{it}}(x_{it+1}|x_{it},u_{it}) : x_{it+1} \in \mathcal{X}_{t+1}\}$ implies that the operator $L_{X_{it+1}|x_{it},U_{it}}$ is invertible. In this case, there is only the tail condition on the function $h(x_{it+1,k},x_{it},u_{it})$ and we can regard h as nonparametric deviation or oscillation from the normal ϕ. Therefore, the invertibility of the integral operator $L_{X_{it+1}|x_{it},U_{it}}$ is appropriate in a nonparametric setting. The condition contains a restriction on the unobsevable and it cannot be verified. A way to justify the condition is invoking the central limit theorem to conclude that $f_{X_{it+1}|X_{it},U_{it}}(x_{it+1}|x_{it},u_{it})$ has an approximate normal distribution and the invertibility permits nontrivial variation around a normal distribution.

In addition, the invertibility of the operator $L_{X_{it+1},x_{it},y_{it-1},X_{it-1}} = L_{X_{it+1}|x_{it},U_{it}} L_{x_{it},y_{it-1},X_{it-1},U_{it}}$ does imply restrictions on the initial condition through the operator $L_{x_{it},y_{it-1},X_{it-1},U_{it}}$. For example, in a case where X_{it} and U_{it} are discrete and the linear operators are matrices, the invertibility of these operators are equivalent to the invertibility of corresponding matrices. However, the operators or matrices may still have a flexible form so that there is no need to specify the initial condition.

Note that when the unobserved component U_{it} is continuous, the invertibility of $L_{X_{it+1}|x_{it},U_{it}}$ implies that the explanatory variables X_{it} contain a continuous element Z_{it}. The existence of the continuous component, Z_{it} is essential. It is impossible to nonparametrically identify a distribution of a continuous unobservable variable only by observed discrete variables. The

X_{it+1}. Honoré and Kyriazidou (2000) and Honoré and Tamer (2006) identify the parameters under certain assumptions on the strictly exogenous covariates.

[14] The choice of ϕ is for simplicity. Please see Hu and Shiu (2011) for general results.

restriction imposed on the continuous Z_{it+1} guarantees that the explanatory variables X_{it+1} contains enough information to identify unobserved component U_{it}. A sufficient condition for identification with continuous U_{it} can be obtained from the well-known completeness property of exponential families.[15] Thus, if $\mathcal{U}_{it}$ is an open set then $\mathcal{X}_{it+1}$ must be an open set.[16] In the case of the intertemporal labor force participation behavior of married women, the covariates X_{it} contain wage and U_{it} includes the unobserved individual skill level or motivation.

Assumption 7.1.4 *(Distinctive eigenvalues) There exists a known function $\omega(\cdot)$ such that $E[\omega(Y_{it})|x_{it}, y_{it-1}, u_{it}]$ is monotonic in u_{it} for any given (x_{it}, y_{it-1}).*

The function $\omega(\cdot)$ may be specified by users, such as $\omega(y) = y$, $\omega(y) = I(y > 0)$, or $\omega(y) = y^2$. For example, we may have $\omega(y) = I(y = 0)$ in the two examples above. In both cases, $E[I(Y_{it} = 0)|x_{it}, y_{it-1}, u_{it}] = F_{\xi_{it}}[-(x'_{it}\beta + \gamma y_{it-1} + u_{it})]$, which is monotonic in u_{it}. Assumption 7.1.4 implies that for all $\hat{U}_{it}, \tilde{U}_{it} \in \mathcal{U}$, the set $\{y : f_{Y_{it}|X_{it},Y_{it-1},\hat{U}_{it}} \neq f_{Y_{it}|X_{it},Y_{it-1},\tilde{U}_{it}}\}$ for any given (x_{it}, y_{it-1}) has a positive probability whenever $\hat{U}_{it} \neq \tilde{U}_{it}$.

Assumption 7.1.5 *(Normalization) For any given $x_{it} \in \mathcal{X}_{it}$, there exists a known functional G such that $G\left[f_{X_{it+1}|X_{it},U_{it}}(\cdot|x_{it}, u_{it})\right] = u_{it}$.*

The functional G may be the mean, the mode, median, or a quantile. For example, we may have $X_{it+1} = X_{it} + U_{it} + h(X_{it})\epsilon_{it}$ with an unknown function $h(\cdot)$ and a zero median independent error ϵ_{it}. Then U_{it} is the median of the density function $f_{(X_{it+1}-X_{it})|X_{it},U_{it}}(\cdot|x_{it}, u_{it})$. The purpose of Assumption 3.5 is to normalize $f_{X_{it+1}|X_{it},U_{it}}$ to be unique in the spectral decomposition and it requires the functional G to map the eigenfunction to a real number. The condition can also be written as $G\left[f_{X_{it+1}|X_{it},U_{it}}(\cdot|x_{it}, u_{it})\right] = l(u_{it})$ for some one-one function $l(\cdot)$ and thus it is not very restrictive.

This assumption imposes a restriction on the covariate evolution. A choice of G depends on how the covariate X_{it} changes over time given the unobserved covariate U_{it}. Hence, observations on the conditional temporal correlation of X_{it} may shed a light on the pick of G. In the case of the intertemporal labor force participation behavior of married women, X_{it} may include annual family income, which often varies with the unobserved time-invariant family characteristics and past economy shock. In this case, setting G as the mode functional seems appropriate.

Main Identification Results

We start our identification with a panel data containing two periods of the dependent variable Y_{it} and three periods of the covariate X_{it}, $\{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}\}$ for $i = 1, 2, ..., n$. The law of total probability leads to

$$f_{X_{it+1},Y_{it},X_{it},Y_{it-1},X_{it-1}} = \int f_{X_{it+1}|Y_{it},X_{it},Y_{it-1},X_{it-1},U_{it}} f_{Y_{it}|X_{it},Y_{it-1},X_{it-1},U_{it}} f_{X_{it},Y_{it-1},X_{it-1},U_{it}} dU_{it},$$

[15] See Newey and Powell (2003) for details.

[16] Assumption 3.3 impose the invertibility of the linear operator $L_{X_{it+1}|x_{it},U_{it}}$ which maps from the domain space $\mathcal{L}^p(\mathcal{U}_t)$ to the range space $\mathcal{L}^p(\mathcal{X}_{t+1})$. The invertibility implies a cardinality relation, the cardinality of $\mathcal{U}_t$ is smaller than the cardinality of $\mathcal{X}_{t+1}$. If U_{it} takes continuous values then X_{it+1} must continuous values.

where we omit the arguments in the density function to make the expressions concise.

Assumption 7.1.1 implies

$$f_{X_{it+1},Y_{it},X_{it},Y_{it-1},X_{it-1}} = \int f_{X_{it+1}|Y_{it},X_{it},Y_{it-1},X_{it-1},U_{it}} f_{Y_{it}|X_{it},Y_{it-1},U_{it}} f_{X_{it},Y_{it-1},X_{it-1},U_{it}} dU_{it}.$$

Then, Assumption 7.1.2 suggests that

$$f_{X_{it+1},Y_{it},X_{it},Y_{it-1},X_{it-1}} = \int f_{X_{it+1}|X_{it},U_{it}} f_{Y_{it}|X_{it},Y_{it-1},U_{it}} f_{X_{it},Y_{it-1},X_{it-1},U_{it}} dU_{it}. \tag{7.13}$$

Based on this equation, we may apply the identification results in Hu and Schennach (2008) to show the all the unknown densities on the RHS are identified from the observed density on the LHS. For any given $(y_{it}, x_{it}, y_{it-1})$, we define operators as follows:

$$\begin{aligned} L_{X_{it+1},y_{it},x_{it},y_{it-1},X_{it-1}} &: \mathcal{L}^p(\mathcal{X}_{t-1}) \to \mathcal{L}^p(\mathcal{X}_{t+1}) \\ (L_{X_{it+1},y_{it},x_{it},y_{it-1},X_{it-1}} h)(u) &= \int f_{X_{it+1},Y_{it},X_{it},Y_{it-1},X_{it-1}}(u, y_{it}, x_{it}, y_{it-1}, x) h(x) dx, \end{aligned}$$

and

$$\begin{aligned} D_{y_{it}|x_{it},y_{it-1},U_{it}} &: \mathcal{L}^p(\mathcal{U}) \to \mathcal{L}^p(\mathcal{U}) \\ (D_{y_{it}|x_{it},y_{it-1},U_{it}} h)(u) &= f_{Y_{it}|X_{it},Y_{it-1},U_{it}}(y_{it}|x_{it}, y_{it-1}, u) h(u). \end{aligned}$$

Similarly, define

$$(L_{x_{it},y_{it-1},X_{it-1},U_{it}} h)(u) = \int f_{X_{it},Y_{it-1},X_{it-1},U_{it}}(x_{it}, y_{it-1}, x, u) h(x) dx.$$

Eq. (7.13) is equivalent to the following operator relationship:

$$L_{X_{it+1},y_{it},x_{it},y_{it-1},X_{it-1}} = L_{X_{it+1}|x_{it},U_{it}} D_{y_{it}|x_{it},y_{it-1},U_{it}} L_{x_{it},y_{it-1},X_{it-1},U_{it}}.$$

Integrating out Y_{it} in Eq. (7.13) leads to $f_{X_{it+1},X_{it},Y_{it-1},X_{it-1}} = \int f_{X_{it+1}|X_{it},U_{it}} f_{X_{it},Y_{it-1},X_{it-1},U_{it}} dU_{it}$, which is equivalent to

$$L_{X_{it+1},x_{it},y_{it-1},X_{it-1}} = L_{X_{it+1}|x_{it},U_{it}} L_{x_{it},y_{it-1},X_{it-1},U_{it}}.$$

with $(L_{X_{it+1},x_{it},y_{it-1},X_{it-1}} h)(u) = \int f_{X_{it+1},X_{it},Y_{it-1},X_{it-1}}(u, x_{it}, y_{it-1}, x) h(x) dx$. We may then apply the spectral decomposition results in Hu and Schennach (2008) to identify $f_{X_{it+1}|X_{it},U_{it}}$, $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$, and $f_{X_{it},Y_{it-1},X_{it-1},U_{it}}$ from $f_{X_{it+1},Y_{it},X_{it},Y_{it-1},X_{it-1}}$. Assumption 7.1.1-7.1.3 enable us to have

$$L_{X_{it+1},y_{it},x_{it},y_{it-1},X_{it-1}} L^{-1}_{X_{it+1},x_{it},y_{it-1},X_{it-1}} = L_{X_{it+1}|x_{it},U_{it}} D_{y_{it}|x_{it},y_{it-1},U_{it}} L^{-1}_{X_{it+1}|x_{it},U_{it}},$$

which implies a spectral decomposition of the observed operators on the LHS. The eigenvalues are the kernel function of the diagonal operator $D_{y_{it}|x_{it},y_{it-1},U_{it}}$ and the eigenfunctions are the kernel function $f_{X_{it+1}|X_{it},U_{it}}$ of the operator $L_{X_{it+1}|x_{it},U_{it}}$. Assumption 7.1.4 make

the eigenvalues distinctive. Since the identification from the spectral decomposition is only identified up to u_{it} and its monotone transformation, we make a normalization assumption, Assumption 7.1.5, to pin down the unobserved covariate u_{it}.

Notice that Theorem 1 in Hu and Schennach (2008) implies that all three densities $f_{X_{it+1}|X_{it},U_{it}}$, $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$, and $f_{X_{it},Y_{it-1},X_{it-1},U_{it}}$ are identified under the assumptions introduced above. The model of interest is described by the density $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$. While the initial condition at period $t-1$ is contained in the joint distribution $f_{X_{it},Y_{it-1},X_{it-1},U_{it}}$, the evolution of the covariates X_{it} is described by $f_{X_{it+1}|X_{it},U_{it}}$.

We summarize our identification results as follows:

Theorem 7.1.1 *Under Assumptions 7.1.1, 7.1.2, 7.1.3, 7.1.4, 7.1.5, the observable joint distribution $f_{X_{it+1},Y_{it},X_{it},Y_{it-1},X_{it-1}}$ uniquely determines the model of interest $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$, together with the evolution density of observed covariates $f_{X_{it+1}|X_{it},U_{it}}$ and the initial joint distribution $f_{X_{it},Y_{it-1},X_{it-1},U_{it}}$.*

The identification procedure is also illustrated in subsection 7.1.7 using a finite dimensional discrete example where the linear operators become matrices. Since the unobserved covariate U_{it} appearing in $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ does not have natural units of measurement or it is unclear which values are appropriate for U_{it}, the partial effects averaged across the distribution of U_{it} are more appealing. The average partial effects are based on the effect on a mean response after averaging the unobserved heterogeneity across the population. Theorem 7.1.1 allows us to obtain the marginal distribution of U_{it},

$$f_{U_{it}} = \int_{\mathcal{X}_{it}} \int_{\mathcal{Y}_{it-1}} \int_{\mathcal{X}_{it-1}} f_{X_{it},Y_{it-1},X_{it-1},U_{it}} dX_{it} dY_{it} dX_{it-1}.$$

Suppose that we are interested in the conditional mean of $\omega(y_t)$, which is a scalar function of y_t. Given (X_{it}, Y_{it-1}) the average structural function (ASF) is defined by

$$ASF(X_{it}, Y_{it-1}) = \int_{\mathcal{U}_{it}} \left[\int_{\mathcal{Y}_{it}} \omega(y_t) f_{Y_{it}|X_{it},Y_{it-1},U_{it}} dY_{it} \right] f_{U_{it}} dU_{it}, \tag{7.14}$$

whose identification has been shown in Corollary 2.1. Then the average partial effect (APE) can be defined by taking derivatives or differences of the above expression (7.14) with respect to elements of (X_{it}, Y_{it-1}). These discussions lead to the following result:

Corollary 7.1.1 *Under Assumptions 7.1.1, 7.1.2, 7.1.3, 7.1.4, 7.1.5, average structural function (ASF) defined in Eq. (7.14) and the average partial effect (APE) can be identified and estimated by a panel data containing two periods of the dependent variable Y_{it} and three periods of the covariate X_{it}, $\{X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1}\}$ for $i = 1, 2, ..., n$.*

Discussion of Assumptions

We discussed the identification assumptions separately in subsection 7.1.3 and now we illustrate these assumptions jointly for the models in Example 1 and Example 2. These

models can be used to describe the following economic behaviors. While Y_{it} denotes the t-th period labor force participation decision and the amount of insurance coverage chosen by an individual for Example 1 and Example 2 respectively, the covariate X_{it} is the non-labor income in both models. Assumptions 7.1.1 allows us to separate the exogenous random shock of the dependent variable in period t, ξ_{it}, from all time-varying error term in the past. It follows that ξ_{it} and U_{it} can be used to decompose the particular error structure in the latent variable formulation of the dependent variable Y_{it}. While ξ_{it} is an exogenous random shock in period t, $U_{it} = V_i + \eta_{it}$ is the sum of the time-invariant heterogeneity and a function of all time-varying variables in the past. This implies that both time-invariant and the past time-varying information are in U_{it}, and the observed (X_{it}, Y_{it-1}) has completely captured the contemporaneous information of Y_{it} other than ξ_{it}. Hence, the present time-varying shocks of labor force participation decision or the amount of insurance coverage are independent of the lagged dependent variables, the non-labor income, and U_{it}.

The definition of U_{it} indicates that conditional on U_{it}, the variation of all past shocks before period t $\{\xi_{i\tau}\}_{\tau<t}$ become trivial.[17] Thus, Assumption 7.1.2 only rules out the immediate effect of the current shock ξ_{it} on the future covariate X_{it+1}. In the economic contexts, the assumption reflects the current exogenous shocks of labor force participation decision or the amount of insurance coverage do not affect the non-labor income in the next period.

The linear independence interpretation for the invertibility of an operator in Hu and Shiu (2011) suggests that the invertibility of $L_{X_{it+1},x_{it},y_{it-1},X_{it-1}}$ in Assumption 7.1.3 can be stated as (1) the family of the joint distributions $\{f_{X_{it+1},X_{it},Y_{it-1},X_{it-1}}(u, x_{it}, y_{it-1}, x) : u \in \widetilde{\mathcal{X}}_{t+1}\}$ where $\widetilde{\mathcal{X}}_{t+1} \subset \mathcal{X}_{t+1}$ has nontrivial variation over the index u in $\widetilde{\mathcal{X}}_{t+1}$ in the function space $\mathcal{L}^p(\mathcal{X}_{t-1})$, and (2) the variation is big enough that every function in $\mathcal{L}^p(\mathcal{X}_{t-1})$ can be approximated by the distributions in the family. The assumption requires some dependence of observed covariates over time. If X_{it} is constant across time, then it violates the condition. In this case, the serially correlated nature of $\mathcal{X}_t$ over time can provide some support of statement (1) but statement (2) is the key assumption to make the invertibility hold.

Next, we discuss the invertibility in Example 1 or the empirical application using the linear independence interpretation. Recall that the dynamic discrete-choice model with an unobserved covariate U_{it}: $Y_{it} = 1\left(X_{it}'\beta + \gamma Y_{it-1} + U_{it} + \xi_{it} \geq 0\right)$ where Y_{it} denotes the t-th period participation decision, and X_{it} is the wage or income variable in that period. First, the invertibility of $L_{X_{it+1},x_{it},y_{it-1},X_{it-1}}$ implies that the conditional distribution of wage or income variables $f_{X_{it+1},X_{it},Y_{it-1},X_{it-1}}(u, x_{it}, y_{it-1}, x)$ over some subset of $\mathcal{X}_{t+1}$ can approximate distributions of wage or income in period $t-1$ well and hence, any income or wage distribution in period $t-1$ has been accounted for by this functional form using the variation in period $t+1$. The independence of income or wage variables over time clearly cause the invertibility to fail. Second, if the unobserved covariate U_{it} contains time-invariant heterogeneity such as motivation or inherent health, the invertibility of $L_{X_{it+1}|x_{it},U_{it}}$ suggests that given the current income variable x_{it} the variation of the functional form $f_{X_{it+1}|X_{it},U_{it}}$ over the future income variables can fully capture the changes or movement of unobserved

[17] Recall $U_{it} = V_i + \eta_{it}$ and $\eta_{it} = \varphi\left(\{X_{i\tau}, Y_{i\tau-1}, \xi_{i\tau}\}_{\tau=0,1,\ldots,t-1}\right)$, η_{it} is a function of all time-varying variables in the past.

motivation or inherent health.

These two models have a point mass at $y = 0$, so we can choose $\omega(y) = I(y = 0)$. Assumption 7.1.4 is automatically satisfied for these limited dependent variable models. Finally, the covariate evolution represents the changing of the non-labor income over time in these models. As mentioned in Assumption 7.1.5, the functional G can be the mean, mode, median, or a quantile. Thus, one of the conditions for Assumption 7.1.5 is that the mode of the distribution of the non-labor income in the next period conditional on the current non-labor income and unobserved covariate u_{it} is equal to the unobserved covariate. Since the unobserved covariate U_{it} contains time-invariant heterogeneity such as motivation or inherent health, it means that the value of non-labor income that occurs most frequently around the location of true level of unobserved motivation or inherent health.

Set $\varepsilon_{it} = \rho\varepsilon_{it-1} + \xi_{it}$ and $\xi_{it} \sim N(0, \sigma_\xi^2)$. Consider the following data generating process (DGP):

$$\begin{aligned} Y_{it} &= g(\beta_0 + \beta_1 X_{it} + \gamma Y_{it-1} + U_{it} + \xi_{it} \geq 0) \quad \text{with} \\ U_{it} &= V_i + \rho\varepsilon_{it-1} \quad \forall \quad i = 1, ..., N; t = 1, ..., T-1, \end{aligned} \tag{7.15}$$

where $g(\cdot)$ can be the 0-1 indicator function or $g(\cdot) = \max(0, \cdot)$ and $V_i \sim N(\mu_v, \sigma_v^2)$. The generating process of covariate evolution has the following form $X_{it+1} = X_{it} + h(X_{it})\epsilon_{it} + U_{it}$ or

$$f_{X_{it+1}|X_{it},U_{it}}(x_{t+1}|x_t, u) = \frac{1}{h(x_t)} f_\epsilon\left(\frac{x_{t+1} - x_t - u}{h(x_t)}\right), \tag{7.16}$$

where f_ϵ is a density function that can be specified under different identification conditions of Assumption 7.1.5.[18] For example, take $f_\epsilon(x) = exp(x - e^x)$ and the mode as the choice of G for Assumption 7.1.5. We will use these settings in the Monte Carlo simulation.

It is straightforward to verify the assumptions with the specific data generating processes except for Assumption 7.1.3. The invertibility of $L_{X_{it+1}|x_{it},U_{it}}$ is equivalent to the completeness of the family $\{f_{X_{it+1}|X_{it},U_{it}}(x_{t+1}|x_t, u) : x_{t+1} \in \mathcal{X}_{it+1}\}$. When $f_\epsilon(x) = exp(x - e^x)$, the covariate evolution belongs to one of exponential families and it is complete by Theorem 2.2 in Newey and Powell (2003). Therefore, $L_{X_{it+1}|x_{it},U_{it}}$ is invertible. Applying the invertibility of $L_{X_{it+1}|x_{it},U_{it}}$ to the integral relation $L_{X_{it+1},x_{it},y_{it-1},X_{it-1}} = L_{X_{it+1}|x_{it},U_{it}} L_{x_{it},y_{it-1},X_{it-1},U_{it}}$ implies that the invertibility of $L_{X_{it+1},x_{it},y_{it-1},X_{it-1}}$ is equivalent to the invertibility of $L_{x_{it},y_{it-1},X_{it-1},U_{it}}$. Utilize Theorem 2.2 in Newey and Powell (2003) again to the family $\{f_{U_{it-1}|X_{it},X_{it-1}} = \frac{1}{h(x_{t-1})} f_\epsilon\left(\frac{-u+x_t-x_{t-1}}{h(x_{t-1})}\right) : u \in \mathcal{U}_{it-1}\}$ for each given x_t and then use it to obtain the completeness of the family $\{f_{X_{it},X_{it-1},U_{it-1}}(x_t, x_{t-1}, u) : u \in \mathcal{U}_{it-1}\}$.[19] Next, pass the completeness of $\{f_{X_{it},X_{it-1},U_{it-1}}(x_t, x_{t-1}, u) : u \in \mathcal{U}_{it-1}\}$ to $\{f_{X_{it},X_{it-1},U_{it}}(x_t, x_{t-1}, u) :$

[18] This generating process is also adopted in Hu and Schennach (2008) and it can be adjusted to a variety of identification conditions, the mean, the mode, median, or a quantile.

[19] Suppose that $h \in \mathcal{L}^p(\mathcal{X}_{it-1})$ and $\int h(x_{t-1}) f_{X_{it},X_{it-1},U_{it-1}}(x_t, x_{t-1}, u) dx_{t-1} = 0$ for any x_t. The equation can be rewritten as $\int h(x_{t-1}) f_{X_{it},X_{it-1}} f_{U_{it-1}|X_{it},X_{it-1}} dx_{t-1} = 0$ for any u_{it-1}. The completeness of $\{f_{U_{it-1}|X_{it},X_{it-1}} : u \in \mathcal{U}_{it-1}\}$ implies that $h(x_{t-1}) f_{X_{it},X_{it-1}} = 0$ and then $h = 0$. We obtain the completeness of the family $\{f_{X_{it},X_{it-1},U_{it-1}}(x_t, x_{t-1}, u) : u \in \mathcal{U}_{it-1}\}$ over $\mathcal{L}^p(\mathcal{X}_{it-1})$.

$u \in \mathcal{U}_{it}\}$ using an integral equation

$$f_{X_{it},X_{it-1},U_{it}} = \int f_{U_{it}|U_{it-1}} f_{X_{it},X_{it-1},U_{it-1}} dU_{it-1}.$$

Since $U_{it} = U_{it-1}$+a normal error, $f_{U_{it}|U_{it-1}}$ is a complete distribution by the normality. We can express the integral equation as an operator relationship and show the operator using $f_{X_{it},X_{it-1},U_{it}}$ as a kernel is invertible. This implies $\{f_{X_{it},X_{it-1},U_{it}}(x_t, x_{t-1}, u) : u \in \mathcal{U}_{it}\}$ is complete and then the family $\{f_{X_{it},Y_{it-1},X_{it-1},U_{it}}(x_t, y_{t-1}, x_{t-1}, u) : u \in \mathcal{U}_{it}\}$ is also complete over $\mathcal{L}^p(\mathcal{X}_{it-1})$. We have reached $L_{X_{it+1},x_{it},y_{it-1},X_{it-1}}$ is invertible.

7.1.4 Estimation

The dynamic panel data model (7.10) specifies the relationship between the dependent variable of interest for an individual i, Y_{it}, and the explanatory variables including a lagged dependent variable Y_{it-1}, a set of possibly time-varying explanatory variables X_{it}, and an unobserved covariate U_{it}. If we are willing to make a normality assumption on ξ_{it}, then the model in Example 1 becomes a probit model and the model in Example 2 becomes a tobit model. The general specification here covers a number of other dynamic nonlinear panel data model in one framework.

Given that the random shocks $\{\xi_{it}\}_{t=0}^{T}$ are exogenous, the conditional distribution $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ is a combination of the function g and the distribution of ξ_{it}. In most applications, the function g and the distribution of ξ_{it} have a parametric form. That means the model may be parameterized in the following form,

$$f_{Y_{it}|X_{it},Y_{it-1},U_{it}}(y_{it}|x_{it}, y_{it-1}, u_{it}; \theta),$$

where θ includes the unknown parameters in both the function g and the distribution of ξ_{it}. Under the rank condition in the regular identification of parametric models, the nonparametric identification of $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ implies that of the parameter θ, and therefore, the identification of the function g and the distribution of ξ_{it}. In general, we may allow $\theta = (b, \lambda)^T$, where b is a finite-dimensional parameter vector of interest and λ is a potentially infinite-dimensional nuisance parameter or nonparametric component.[20] What is not specified in the model is the evolution of the covariate X_{it}, together with the unobserved component U_{it}, i.e., $f_{X_{it+1}|X_{it},U_{it}}$, and the initial joint distribution of all the variables $f_{X_{it},Y_{it-1},X_{it-1},U_{it}}$. We consider the nonparametric elements $(f_{X_{it+1}|X_{it},U_{it}},\ \lambda,\ f_{X_{it},Y_{it-1},X_{it-1},U_{it}})^T$ as infinite-dimensional nuisance parameters in our semi-parametric estimator.

Our semi-parametric sieve MLE does not require the initial condition assumption for the widely used panel data models, such as dynamic discrete-response models and dynamic censored models. In Section 7.1.3, we have shown equation (7.13) uniquely determines $(f_{X_{it+1}|X_{it},U_{it}},\ f_{Y_{it}|X_{it},Y_{it-1},U_{it}},\ f_{X_{it},Y_{it-1},X_{it-1},U_{it}})^T$. While the dynamic panel data model component $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ will be parameterized, the other components are treated as non-

[20] A partition of θ into finite-dimensional parameters and infinite-dimensional parameters does not affect our sieve MLE. More examples of a partition can be found in Shen (1997).

parametric nuisance functions. Equation (7.13) implies

$$\begin{aligned}\alpha_0 &\equiv (f_{X_{it+1}|X_{it},U_{it}}, \theta_0, f_{X_{it},Y_{it-1},X_{it-1},U_{it}})^T \\ &= \arg \max_{(f_1,\theta,f_2)^T \in \mathcal{A}} E \ln \int f_1(x_{it+1}|x_{it},u_{it}) f_{Y_{it}|X_{it},Y_{it-1},U_{it}}(y_{it}|x_{it},y_{it-1},u_{it};\theta) \\ &\qquad \times f_2(x_{it},y_{it-1},x_{it-1},u_{it}) du_{it},\end{aligned}$$

which suggests a corresponding semi-parametric sieve MLE using an i.i.d. sample $\left\{x_{it+1}, y_{it}, x_{it}, y_{it-1}, x_{it-1}\right\}_{i=1}^n$,

$$\begin{aligned}\widehat{\alpha}_n &\equiv \left(\hat{f}_1, \hat{\theta}, \hat{f}_2\right)^T \\ &= \arg \max_{(f_1,\theta,f_2)^T \in \mathcal{A}_n} \frac{1}{n} \sum_{i=1}^n \ln \int f_1(x_{it+1}|x_{it},u_{it}) f_{Y_{it}|X_{it},Y_{it-1},U_{it}}(y_{it}|x_{it},y_{it-1},u_{it};\theta) \\ &\qquad \times f_2(x_{it},y_{it-1},x_{it-1},u_{it}) du_{it}.\end{aligned} \tag{7.17}$$

The function space $\mathcal{A}$ contains the corresponding true densities and $\mathcal{A}_n$ is a sequence of approximating sieve spaces.

Our estimator is a direct application of the general semi-parametric sieve MLE in Shen (1997), Chen and Shen (1998), and Ai and Chen (2003). In the appendix, we provide sufficient conditions for the consistency of our semi-parametric estimator $\widehat{\alpha}_n$ and those for the $\sqrt{n}$ asymptotic normality of the parametric component $\widehat{b}$. The asymptotic theory of the proposed sieve MLE and the detailed development of sieve approximations of the nonparametric components are also provided in Online Appendix.

With the consistency of the semi-parametric estimator $\widehat{\alpha}_n$, a consistent estimator of the average structural function (ASF) can be obtained by

$$ASF(X_t, Y_{t-1}) = \int_{\mathcal{U}_t} \left[\int_{\mathcal{Y}_t} \omega(y_t) f_{Y_t|X_t,Y_{t-1},U_t}(y_t|x_t,y_{t-1},u_t;\hat{\theta}) dY_t \right] \hat{f}_2(u_t) du_t, \tag{7.18}$$

where $\hat{f}_2(U_t) = \int_{\mathcal{X}_t} \int_{\mathcal{Y}_{t-1}} \int_{\mathcal{X}_{t-1}} \hat{f}_2(X_t, Y_{t-1}, X_{t-1}, U_t) dX_t dY_{t-1} dX_{t-1}$. Thus, the average partial effects of the state dependence at interesting values of the explanatory variables can be computed by changes or derivatives of equation (7.18) with respect to Y_{t-1}.

Note that the proposed sieve MLE only needs 3 periods. This means that when a DGP is generated through the dynamic process (7.10), three-periods data are enough to recovery the parameter of the interest θ. When there are more periods of data, the approach is still tractable. For example, if $T = 4$ and we assume the dynamic panel data specification (7.10), then estimation results from periods $1, 2$, and 3 should be the same as ones from $2, 3$, and 4. If the estimated results are significantly different, we would suspect model misspecification. Under the assumptions of stationary and ergodicity, an alternative way to deal with data more than 3 periods is to transform the data into 3 periods of data by rearranging them as 3 periods of data and stacking them into a larger cross-sectional data. For example, suppose that there are 5 periods of data $\{D_t, D_{t+1}, D_{t+2}, D_{t+3}, D_{t+4}\}$. It can be transformed into

three observations of three periods of data, i.e., $\{D_t, D_{t+1}, D_{t+2}\}$, $\{D_{t+1}, D_{t+2}, D_{t+3}\}$, and $\{D_{t+2}, D_{t+3}, D_{t+4}\}$.

For a model with a larger number of observed covariates, we can consider a single-index response model with $X_{it}'\beta$. That is: X_{it} is a d-dimensional vector of explanatory variables, $X_{it}'\beta$ is the index, the scalar product of X_{it} with β, a vector of parameters whose values are unknown. Since our assumptions do not exclude time dependence in covariates, time dummies are allowed to be in X_{it}. Many widely used parametric models have this form. In our empirical application, we adopt this approach to deal with a case of many observed covariates. The part (ii) of Assumption B.4. requires that $k_{ni}/n \to 0$ for $i = 1, \lambda, 2$. Thus, the rate of convergence depends on the degree of the sieve approximations since higher degree of sieve spaces provide better approximations. When X_{it} is a d-dimensional vector and the index form is not used, the degrees of approximation has to be increased proportionally in order to get better approximation of these nuisance component, $f_1(x_{t+1}|x_t, u_t; \delta_1)$ and $f_2(x_t, y_{t-1}, x_{t-1}, u_t; \delta_2)$. It follows that the larger the dimension of X_{it}, the slower the rate of convergence. Thus, the curse of dimensionality may be an issue if researchers are interested in the nuisance component $f_1(x_{t+1}|x_t, u_t; \delta_1)$ and $f_2(x_t, y_{t-1}, x_{t-1}, u_t; \delta_2)$, but the convergence speed of the parametric part is still root-n.

Implementation

As we discussed above, we propose a semi-parametric sieve MLE using an i.i.d. sample $\{x_{it+1}, y_{it}, x_{it}, y_{it-1}, x_{it-1}\}$ for $i = 1, 2, ..., n$. The unknown densities are associated with the observed distribution as follows:

$$f_{X_{it+1},Y_{it},X_{it},Y_{it-1},X_{it-1}} = \int f_{X_{it+1}|X_{it},U_{it}} f_{Y_{it}|X_{it},Y_{it-1},U_{it}} f_{X_{it},Y_{it-1},X_{it-1},U_{it}} dU_{it}.$$

The parametric part is the model of interest $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}(y_{it}|x_{it}, y_{it-1}, u_{it}; \theta)$. The two nonparametric nuisance functions include $f_{X_{it+1}|X_{it},U_{it}}$ and $f_{X_{it},Y_{it-1},X_{it-1},U_{it}}$. The sieve MLE transforms a semi-parametric MLE to a parametric MLE by replacing the nonparametric nuisance functions with their Fourier approximations. For example, the sieve estimator for the covariate evolution may be constructed by the Fourier series as follows:

$$f_1(x_{t+1}|x_t, u_t; \delta_1) = \sum_{i=0}^{i_n} \sum_{j=0}^{j_n} \sum_{k=0}^{k_n} \delta_{1,ijk} \varphi_{1i}(x_{t+1} - u_t) \varphi_{2j}(x_t) \varphi_{3k}(u_t),$$

where i_n, j_n, k_n are smoothing parameters and $\varphi_{1i}, \varphi_{2j}, \varphi_{3k}$ are known basis functions. Similarly, we may have a sieve approximation of the initial joint density, $f_2(x_{it}, y_{it-1}, x_{it-1}, u_{it}; \delta_2)$, where δ_2 is a vector of all the sieve coefficients. The fact that the parametric functions $f_1(x_{t+1}|x_t, u_t; \delta_1)$ and $f_2(x_{it}, y_{it-1}, x_{it-1}, u_{it}; \delta_2)$ are approximations of probability density functions implies certain restrictions on the sieve coefficients (δ_1, δ_2), which is discussed in Online Appendix. In the sieve MLE, we may estimate $(\theta, \delta_1, \delta_2)$ as a parametric MLE with

a density function as follows:

$$f\left(x_{it+1}, y_{it}, x_{it}, y_{it-1}, x_{it-1}; \theta, \delta_1, \delta_2\right) = \int f_1(x_{it+1}|x_{it}, u_t; \delta_1) f_{Y_{it}|X_{it},Y_{it-1},U_{it}}(y_{it}|x_{it}, y_{it-1}, u_{it}; \theta) \times f_2(x_{it}, y_{it-1}, x_{it-1}, u_{it}; \delta_2) du_{it}.$$

In Online Appendix, we show the consistency and asymptotic normality as sample size goes to infinity.

7.1.5 Monte Carlo Evidence

In this section we present a Monte Carlo study that investigates the finite sample properties of the proposed sieve MLE estimators in the two different settings, dynamic discrete choice models and dynamic censored models. We start with the specification of the models as follows.

Semi-parametric Dynamic Probit Models

First, we adopt a parametric assumption for ε_{it}. Suppose that ε_{it} has a stationary AR(1) with an independent Gaussian white noise process, $\varepsilon_{it} = \rho\varepsilon_{it-1} + \xi_{it}$, $\xi_{it} \sim N(0, 1/2)$. Denote $\Phi_{\xi_{it}}$ and $\phi_{\xi_{it}}$ as the CDF and PDF of the independent error ξ_{it}, respectively. We have

$$f_{Y_{it}|X_{it},Y_{it-1},U_{it}} = \Phi_{\xi_{it}}\left(X_{it}'\beta + \gamma Y_{it-1} + U_{it}\right)^{Y_{it}} \left[1 - \Phi_{\xi_{it}}\left(X_{it}'\beta + \gamma Y_{it-1} + U_{it}\right)\right]^{1-Y_{it}},$$

with $U_{it} = V_i + \rho\varepsilon_{it-1}$.

The density $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ is fully parameterized and θ only contain the parametric component $b = (\gamma, \beta)^T$. We approximate $f_{X_{it+1}|X_{it},U_{it}}$, and $f_{X_{it},Y_{it-1},X_{it-1},U_{it}}$ by truncated series in the estimation. The estimator of average structural function (ASF) in the dynamic probit model is

$$ASF(X_t, Y_{t-1}) = \int_{\mathcal{U}_t} \Phi_{\xi_{it}}\left(X_t'\beta + \gamma Y_{t-1} + U_t\right) f_2(U_t) dU_t, \tag{7.19}$$

which represents the conditional mean of $\omega\left(y_t\right) = y_t$.

Semi-parametric Dynamic Tobit Models:

We also assume that ε_{it} has a stationary AR(1) with an independent Gaussian white noise process, $\varepsilon_{it} = \rho\varepsilon_{it-1} + \xi_{it}$. This gives

$$f_{Y_{it}|X_{it},Y_{it-1},U_{it}} = \left[1 - \Phi_{\xi_{it}}\left(X_{it}'\beta + \gamma Y_{it-1} + U_{it}\right)\right]^{\mathbf{1}(Y_{it}=0)} \phi_{\xi_{it}}(y_{it} - X_{it}'\beta - \gamma Y_{it-1} - U_{it})^{\mathbf{1}(Y_{it}>0)} \tag{7.20}$$

$$= \left[1 - \Phi\left(\frac{X_{it}'\beta + \gamma Y_{it-1} + U_{it}}{\sigma_\xi}\right)\right]^{\mathbf{1}(Y_{it}=0)} \times \left[\frac{1}{\sigma_\xi}\phi\left(\frac{y_{it} - X_{it}'\beta - \gamma Y_{it-1} - U_{it}}{\sigma_\xi}\right)\right]^{\mathbf{1}(Y_{it}>0)},$$

and the parameter is $\theta = b = (\gamma, \beta, \sigma_\xi^2)^T$. Since $\xi_{it} \sim N(0, \sigma_\xi)$, $E_{Y_t}[y_t|X_t, Y_{t-1}, U_t] = \Phi\left(\frac{X_t'\beta+\gamma Y_{t-1}+U_t}{\sigma_\xi}\right)(X_t'\beta + \gamma Y_{t-1} + U_t) + \sigma_\xi \phi\left(\frac{X_t'\beta+\gamma Y_{t-1}+U_t}{\sigma_\xi}\right)$. The estimator of ASF in the dynamic tobit model is

$$ASF(X_t, Y_{t-1}) = \int_{\mathcal{U}_t} \left[\Phi\left(\frac{X_t'\beta + \gamma Y_{t-1} + U_t}{\sigma_\xi}\right)(X_t'\beta + \gamma Y_{t-1} + U_t) + \sigma_\xi \phi\left(\frac{X_t'\beta + \gamma Y_{t-1} + U_t}{\sigma_\xi}\right)\right] f_2(U_t) dU_t. \tag{7.21}$$

The data generating process for dynamic discrete choice models and dynamic censored models in the Monte Carlo experiments are according to the following processes respectively:

$$\begin{aligned} Y_{it} &= 1(\beta_0 + \beta_1 X_{it} + \gamma Y_{it-1} + U_{it} + \xi_{it} \geq 0) \quad \text{with} \\ U_{it} &= V_i + \rho \varepsilon_{it-1} \quad \forall \quad i = 1, ..., N; t = 1, ..., T-1. \end{aligned} \tag{7.22}$$

and

$$\begin{aligned} Y_{it} &= \max\{\beta_0 + \beta_1 X_{it} + \gamma Y_{it-1} + U_{it} + \xi_{it}, 0\} \quad \text{with} \\ U_{it} &= V_i + \rho \varepsilon_{it-1} \quad \forall \quad i = 1, ..., N; t = 1, ..., T-1. \end{aligned} \tag{7.23}$$

where $V_i \sim N(1, 1/2)$. To construct the sieve MLE, it is necessary to integrate out the unobserved covariate U_{it}. Here U_{it} has an unbounded domain $(-\infty, \infty)$ and we adopted Gauss-Hermite quadrature for approximating the value of the integral. We consider the mode condition for Assumption 7.1.5, and use $f_\epsilon(x) = exp(x - e^x)$ in equation (7.16) for all simulated data. In addition, we set $h(x) = 0.3\exp(-x)$ to allow heterogeneity and assume the initial observation (y_0, x_0) and the initial component ξ_0 $(=\epsilon_{i0})$ equal to zero. As discussed in subsection 3.3, these data generating processes satisfy the identification Assumptions 7.1.1-7.1.5.

We consider five different values of $(\gamma, \sigma_\xi^2, \rho)$ in the experiments: $(\gamma, \sigma_\xi^2, \rho)$ = (0,0.5,0), (0,0.5,0.5), (1,0.5,0), (1,0.5,0.5), (1,0.5,-0.5) and the parameters of the intercept and the exogenous variable are held fixed: $\beta_0 = 0$ and $\beta_1 = -1$. In summary, the data generating processes are as follows:

$$\begin{aligned} \text{DGP I:} \quad & (\beta_0, \beta_1, \gamma, \sigma_\xi^2, \rho) = (0, -1, 0, 0.5, 0) \\ \text{DGP II:} \quad & (\beta_0, \beta_1, \gamma, \sigma_\xi^2, \rho) = (0, -1, 0, 0.5, 0.5) \\ \text{DGP III:} \quad & (\beta_0, \beta_1, \gamma, \sigma_\xi^2, \rho) = (0, -1, 1, 0.5, 0) \\ \text{DGP IV:} \quad & (\beta_0, \beta_1, \gamma, \sigma_\xi^2, \rho) = (0, -1, 1, 0.5, 0.5) \\ \text{DGP V:} \quad & (\beta_0, \beta_1, \gamma, \sigma_\xi^2, \rho) = (0, -1, 1, 0.5, -0.5). \end{aligned}$$

The first two DGPs are not state dependence ($\gamma = 0$) while the rest are state dependent with $\gamma = 1$. A sample size N=500 is considered.[21] To secure a more stationary sample, the

[21] Simulation results for other two different sample sizes, N=250, 1000 are online.

sampling data are drawn over $T = 7$ periods but only last three periods are utilized. 100 simulation replications are conducted at each estimation.

Table 7.1 presents simulation results under the semi-parametric probit model. The simulation results of DGP I (only allows for unobserved heterogeneity) show small standard deviations exist in the structural model coefficients (β_0, γ) comparing to the benchmark estimator. For DGP II, the results have downward bias in the structural model coefficient β_1. In addition, with nontrivial transitory component ($\rho \neq 0$) in DGP II, the standard deviations of $(\beta_0, \beta_1, \gamma)$ are not much different from DGP I. As for DGPs with nontrivial state dependence, bias for $(\beta_0, \beta_1, \gamma)$ for these DGPs is around 0.01 or less and their standard deviations are around 0.1. The coefficient estimators of γ in these DGPs have very small bias for all sample sizes, which means that our estimation for state dependence is very precise among processes with serial correlation ($\rho \neq 0$). In general, the means and medians of (β_1, γ) are very close to each other, reflecting little skewness in their respective distributions. Table 7.2 shows the simulation of the average partial effects in dynamic probit models in these DGPs. When there is no state dependence (DGP I & II), the estimates for average partial effects do not vary much with the lagged value Y_{t-1}. However, when DGPs contain state dependence, the difference in the average responses are up to 0.12. Results using the benchmark estimator have much larger standard deviations than ones using the proposed estimator.

Table 7.3 reports the results of estimates for the semi-parametric tobit model. In the tobit model, there is negative bias in β_1 for all DGPs. In tobit case, we have additional parameters to estimate, σ_ξ^2. There is upward bias of the parameter in all DGPs and their standard deviations are a little bit higher in DGPs with nontrivial state dependence. For these DGPs with positive state dependence, estimation results of γ show that there is small bias and precision is within 0.05. Also, the means and medians of all model parameters are not much different, reflecting low degree of skewness in distributions. Table 7.4 shows the results of the average partial effects in dynamic tobit models. There are larger standard deviations of average structural functions and state dependence in DGPs with positive state dependence. Similar to the results in Table 7.2, results using the benchmark estimator have much larger standard deviations than ones in the proposed estimator.

In some estimation results of parameters, the simulation standard deviation is smaller for the proposed semi-parametric estimator than for the benchmark parametric MLE. An explanation for this observation is that we have adopted Gauss-Hermite quadrature for approximating the value of the integral in the sieve MLE and the distribution of the weights of Gauss-Hermite quadrature are close to a normal distribution. On the other hand, in our simulation design, the unobserved covariate U_{it} is normally distributed. This may reduce the simulated standard deviation because in this case the weight function used in numerical integration has the same functional form as a normal PDF.

There are two nuisance parameters, $f_{X_{t+1}|X_t,U_t}$ and $f_{X_t,Y_{t-1},X_{t-1},U_t}$ in our Monte Carlo simulation and we use Fourier series to approximate the evolution density and the square root of the initial joint distribution. Since a higher dimensional sieve space is constructed by tensor product of univariate sieve series, approximation series can be formed from several

univariate Fourier series. In the semi-parametric probit model, while in the approximation of the evolution densities we use three univariate Fourier series with the number of term, $i_n = 5$, $j_n = 2$, and $k_n = 2$, in the approximation of the initial joint distribution we have $i_n = 5$, $j_n = 2$, $k_n = 2$, and $l_n = 2$.[22] While a formal selection rule for these smoothing parameters would be desirable, it is difficult to provide a general guideline. From our experience, the estimation of the finite-dimensional parameters θ is not very sensitive to these smoothing parameters. If one cares about estimation of nonparametric density functions, one should pick the smoothing parameters to minimize the approximate mean squared errors of the estimator. In the Monte Carlo study, this is relatively easy to do because the true values are known. But in empirical applications where the true values of the parameters are unknown, it is still a difficult task. A rule of thumb is to pick the smoothing parameters such that the estimates are not sensitive to small variations in the smoothing parameter.[23] While the Fourier approximations to the evolution density $f_{X_{t+1}|X_t,U_t}$ have the density restriction and the identification restriction, there exists only a density restriction for the approximations to the square root of the initial joint distribution $f^{1/2}_{X_t,Y_{t-1},X_{t-1},U_t}$ using Fourier basis.[24] The semi-parametric sieve MLE using this construction does not encounter any negative integral inside the logarithm on equation (7.17) in our Monte Carlo study. As for the semi-parametric tobit model, we have similar choices of approximation series. The detailed sieve expression of these nuisance parameters can be found in Online Appendix.

The standard deviations can be computed from bootstraps from draws of the original sample. The use of nonparametric bootstrap provides an asymptotically valid standard deviations for the sieve MLE estimate for the finite dimensional parameters θ. The discussion of the consistency of the ordinary nonparametric bootstrap for θ can be found in Chen et al. (2003). Set $Z_{ti} = (X_{it+1}, Y_{it}, X_{it}, Y_{it-1}, X_{it-1})$ and then define a moment function as

[22]The numbers of term, i_n, j_n, and k_n represent the length of three univariate Fourier series. See Online Appendix for details.

[23]There is no justified general rule on the choice of number of series terms. For each smoothing parameter, a minimum choice of number of terms is 2 because a sieve series with each smoothing parameter less than 2 is too restrictive and may not approximate well. Thus, each smoothing parameter should be at least 2. Start with an approximation series whose smoothing parameter is 2 in each univariate series and construct a corresponding likelihood to conduct Monte Carlo experiment. If the result of the simulation based on the approximation series is not satisfactory, then try to add more terms. In this case, we added more terms in p_{1i} and q_i while fixing other univariate series because it is easier to add terms in one particular univariate series without changing the whole structure of the approximation series. If this does not work well then do the adding and fixing step to other univariate series. The process can continue to an approximation series whose smoothing parameter is at least 3 in each univariate series. Therefore, the search procedure is complete and help us determine the number of series terms. In addition, a discussion in Hu and Schennach (2008) suggests that a suitable choice of the smoothing parameters lies between short series and long series where the smoothing bias and the statistical noise dominate respectively.

[24]An approximation series to a positive density function may take negative values. A natural log value of a negative value is infinity and this may make the construction of log likelihood function infeasible. Using an approximation series to the square root of the initial joint distribution yields an positive approximation to the positive density function.

$m(Z_t, \theta, f_1, f_2) = \ln f\left(x_{t+1}, y_t, x_t, y_{t-1}, x_{t-1}; \theta, \delta_1, \delta_2\right)$, where

$$f\left(x_{t+1}, y_t, x_t, y_{t-1}, x_{t-1}; \theta, \delta_1, \delta_2\right) = \int f_1(x_{t+1}|x_t, u_t; \delta_1) f_{Y_t|X_t, Y_{t-1}, U_t}(y_t|x_t, y_{t-1}, u_t; \theta) \times f_2(x_t, y_{t-1}, x_{t-1}, u_t; \delta_2) du_t.$$

The notation connects the proposed sieve MLE to the setting in Chen et al. (2003). Sufficient conditions for the bootstrap validity in Chen et al. (2003) include the identification of a parameter, the approximation of a sequence of sieve spaces to infinite dimensional parameters, and the regularity conditions of the moment function. These conditions are close to conditions of the consistency and asymptotic normality in the Appendix B.[25] In a sieve related estimation method, Ai and Chen (2003) also adopted bootstrap standard deviations as standard deviations of their sieve minimum distance estimator in the simulation study.

In summary, the Monte Carlo study shows that our semi-parametric sieve MLE performs well with a finite sample since mean and median estimates are close to the true values with reasonable standard deviations.

[25] For example, we have Assumption B.5 for that $\ln f_{Z_t}(z_t; \alpha)$ is Hölder continuous and Chen et al. (2003) provided Hölder continuity as one of primitive sufficient assumptions for their bootstrap result. Therefore, we may not need to impose extra assumptions on the validity of bootstrapping standard errors.

Table 7.1: Monte Carlo Simulation of Semi-parametric Probit model (n=500)

		Parameters		
	DGP	β_0	β_1	γ
DGP I:	true value	0	-1	0
	mean benchmark	-0.033	-1.011	0.059
	median benchmark	0.017	-1.016	0.008
	standard deviation	0.387	0.065	0.452
	mean estimate	0.008	-0.994	-0.013
	median estimate	0.010	-1.006	-0.002
	standard deviation	0.086	0.103	0.108
DGP II:	true value	0	-1	0
	mean benchmark	0.015	-1.013	0.024
	median benchmark	0.006	-1.011	0.021
	standard deviation	0.125	0.065	0.101
	mean estimate	-0.003	-1.010	0.007
	median estimate	-0.012	-1.004	0.011
	standard deviation	0.087	0.095	0.110
DGP III:	true value	0	-1	1
	mean benchmark	0.002	-1.004	0.998
	median benchmark	-0.001	-1.005	0.997
	standard deviation	0.134	0.071	0.093
	mean estimate	0.008	-0.991	0.997
	median estimate	0.016	-0.994	1.000
	standard deviation	0.093	0.105	0.106
DGP IV:	true value	0	-1	1
	mean benchmark	-0.052	-0.999	1.056
	median benchmark	-0.014	-1.000	1.015
	standard deviation	0.412	0.055	0.411
	mean estimate	-0.005	-1.005	1.008
	median estimate	0.003	-1.024	1.010
	standard deviation	0.092	0.104	0.121
DGP V:	true value	0	-1	1
	mean benchmark	0.012	-1.010	1.000
	median benchmark	0.001	-1.011	1.001
	standard deviation	0.112	0.066	0.096
	mean estimate	-0.001	-0.996	0.996
	median estimate	0.012	-1.002	0.982
	standard deviation	0.112	0.095	0.093

Note: The simulated date has 7 periods but only last 3 periods are used to construct the sieve MLE in the semi-parametric probit model. The benchmark estimator is an unfeasible MLE using the unobserved covariate U_{it}. Standard deviations of the parameters are computed by the standard deviation of the estimates across 100 simulations and called (simulation) standard deviations.

Table 7.2: Simulation of Average Structural Functions in Probit model (n=500)

	State Dependence	Average Structural Functions	
DGP I:	$Y_{t-1}=0$	mean benchmark	0.281
		standard deviation	(0.214)
		mean estimate	0.574
		standard deviation	(0.029)
	$Y_{t-1}=1$	mean benchmark	0.281
		standard deviation	(0.214)
		mean estimate	0.572
		standard deviation	(0.035)
DGP II:	$Y_{t-1}=0$	mean benchmark	0.307
		standard deviation	(0.216)
		mean estimate	0.582
		standard deviation	(0.029)
	$Y_{t-1}=1$	mean benchmark	0.307
		standard deviation	(0.216)
		mean estimate	0.580
		standard deviation	(0.034)
DGP III:	$Y_{t-1}=0$	mean benchmark	0.301
		standard deviation	(0.219)
		mean estimate	0.572
		standard deviation	(0.021)
	$Y_{t-1}=1$	mean benchmark	0.640
		standard deviation	(0.204)
		mean estimate	0.696
		standard deviation	(0.028)
DGP IV:	$Y_{t-1}=0$	mean benchmark	0.265
		standard deviation	(0.220)
		mean estimate	0.584
		standard deviation	(0.036)
	$Y_{t-1}=1$	mean benchmark	0.587
		standard deviation	(0.233)
		mean estimate	0.707
		standard deviation	(0.045)
DGP V:	$Y_{t-1}=0$	mean benchmark	0.282
		standard deviation	(0.203)
		mean estimate	0.586
		standard deviation	(0.036)
	$Y_{t-1}=1$	mean benchmark	0.614
		standard deviation	(0.218)
		mean estimate	0.717
		standard deviation	(0.048)

Note: The average structural functions are reported at the mean value of the explanatory variable and two different outcomes of Y_{t-1}, 0 and 1. Standard deviations of these average structural functions are computed by the standard deviation of the estimates across 100 simulations and called (simulation) standard deviations. The true values of ASF are computed using the unobserved covariate U_{it}. Average partial effects of Y_{t-1} can be obtained by taking differences of average structural functions at $Y_{t-1}=0$, and $Y_{t-1}=1$.

Table 7.3: Monte Carlo Simulation of Semi-parametric Tobit model (n=500)

		Parameters			
	DGP	β_0	β_1	γ	σ_ξ^2
DGP I:	true value	0	-1	0	0.5
	mean benchmark	0.001	-1.002	-0.023	0.502
	median benchmark	-0.003	-1.003	0.002	0.502
	standard deviation	0.103	0.064	0.289	0.084
	mean estimate	0.007	-1.006	0.002	0.525
	median estimate	0.006	-0.992	0.009	0.523
	standard deviation	0.092	0.111	0.103	0.031
DGP II:	true value	0	-1	0	0.5
	mean benchmark	-0.013	-0.994	-0.009	0.494
	median benchmark	-0.003	-0.991	-0.009	0.496
	standard deviation	0.088	0.049	0.128	0.065
	mean estimate	0.001	-1.009	0.017	0.526
	median estimate	-0.014	-1.009	0.019	0.524
	standard deviation	0.112	0.096	0.098	0.030
DGP III:	true value	0	-1	1	0.5
	mean benchmark	0.001	-1.004	1.002	0.499
	median benchmark	0.004	-1.000	1.000	0.500
	standard deviation	0.096	0.057	0.052	0.052
	mean estimate	0.015	-1.011	0.989	0.528
	median estimate	0.014	-1.003	0.994	0.526
	standard deviation	0.100	0.112	0.114	0.035
DGP IV:	true value	0	-1	1	0.5
	mean benchmark	-0.001	-1.006	1.004	0.501
	median benchmark	-0.002	-1.013	1.001	0.506
	standard deviation	0.084	0.056	0.047	0.051
	mean estimate	0.007	-1.015	0.988	0.501
	median estimate	0.017	-1.023	0.986	0.523
	standard deviation	0.093	0.103	0.101	0.036
DGP V:	true value	0	-1	1	0.5
	mean benchmark	0.001	-1.007	1.005	0.502
	median benchmark	0.003	-1.003	1.007	0.505
	standard deviation	0.072	0.045	0.057	0.055
	mean estimate	-0.002	-1.030	0.997	0.528
	median estimate	0.008	-1.026	0.996	0.527
	standard deviation	0.108	0.099	0.120	0.035

Note: The simulated date has 7 periods but only last 3 periods are used to construct the sieve MLE in the semi-parametric Tobit models. The benchmark estimator is an unfeasible MLE using the unobserved covariate U_{it}. Standard deviations of the parameters are computed by the standard deviation of the estimates across 100 simulations and called (simulation) standard deviations.

Table 7.4: Simulation of Average Effects in Tobit model (n=500)

	Average Structural Functions		State Dependence	
DGP I:	mean benchmark	0.171	mean benchmark	0.314
	standard deviation	(0.243)	standard deviation	(0.253)
	mean estimate	0.357	mean estimate	0.399
	standard deviation	(0.080)	standard deviation	(0.069)
DGP II:	mean benchmark	0.131	mean benchmark	0.263
	standard deviation	(0.164)	standard deviation	(0.210)
	mean estimate	0.360	mean estimate	0.407
	standard deviation	(0.096)	standard deviation	(0.086)
DGP III:	mean benchmark	0.474	mean benchmark	1.189
	standard deviation	(0.353)	standard deviation	(0.519)
	mean estimate	0.620	mean estimate	1.016
	standard deviation	(0.146)	standard deviation	(0.187)
DGP IV:	mean benchmark	0.437	mean benchmark	1.116
	standard deviation	(0.353)	standard deviation	(0.537)
	mean estimate	0.652	mean estimate	1.045
	standard deviation	(0.106)	standard deviation	(0.143)
DGP V:	mean benchmark	0.468	mean benchmark	1.159
	standard deviation	(0.361)	standard deviation	(0.535)
	mean estimate	0.655	mean estimate	1.073
	standard deviation	(0.149)	standard deviation	(0.180)

Note: The average structural functions are reported at the mean value of the explanatory variable including the lagged dependent variable. Standard deviations of these estimation results are computed by the standard deviation of the estimates across 100 simulations and called (simulation) standard deviations. Average partial effects of Y_{t-1} or State Dependence can be obtained by taking the derivative of ASF at means. The true values of ASF and State Dependence are computed using the unobserved covariate U_{it}.

7.1.6 Empirical Example

In this section, we apply our estimator to a dynamic discrete choice model, which describes the labor force participation decisions of married women given their past participation state and other covariates. The advantage of our estimator is that our model may include (i) arbitrary and unspecified correlated random effects between unobserved time-invariant factors such as skill level or motivation and time-varying X_{it}'s, and (ii) we require no initial conditions assumption.[26] Hyslop (1999) also studied a similar empirical model with less general assumptions but specified parametric forms of the unobserved heterogeneity V_i and AR(1) time dependence ρ of the transitory error component ε_{it}. Since these two terms are

[26] In Hyslop (1999), a correlated random-effects (CRE) specification for v_i is:

$$v_i = \sum_{s=0}^{T} (\delta_{1s} \cdot (\#\text{Kids0-2})_{is} + \delta_{2s} \cdot (\#\text{Kids3-5})_{is} + \delta_{3s} \cdot (\#\text{Kids6-17})_{is}) + \sum_{s=0}^{T-1} \delta_{4s} \cdot y_{mtis} + \eta_i,$$

where y_{mtis} is i's transitory nonlabor income in year s. An alternative CRE specification can be:

$$v_i = \delta_1 \cdot (\overline{\#\text{Kids0-2}})_i + \delta_2 \cdot (\overline{\#\text{Kids3-5}})_i + \delta_3 \cdot (\overline{\#\text{Kids6-17}})_i + \delta_4 \cdot \bar{y}_{mti} + \eta_i,$$

where $\bar{x}_i = \sum_{t=0}^{T} x_{it}$.

not separately identified from our main result Theorem 7.1.1, the empirical study here will focus on the parameters of exogenous explanatory variables and lagged dependent variable not the distributions of the error terms. On the other hand, these estimations might not be comparable across specifications, because of the estimator-specific normalizations in binary choice models. Since the average partial effect is identified in Corollary 7.1.1, the empirical study also focuses on comparable average partial effects.

Specifications and Estimation Results

According to a theoretical model in Hyslop (1999), the labor force participation decisions of married women depend on whether or not their market wage offer exceeds their reservation wage, which in turn may depend on their past participation state. Suppose Y_t is the t-th period participation decision, W_t is the wage, and W_{0t}^* is a reservation wage. Then period t participation decision can be formulated by

$$Y_t = 1(W_t > W_{0t}^* - \gamma Y_{t-1}) \tag{7.24}$$

where $1(\cdot)$ denotes an indicator function that is equal to 1 if the expression is true and 0 otherwise. An empirical reduced form specification for equation (7.24) is the following

$$Y_{it} = 1(X_{it}'\beta + \gamma Y_{t-1} + U_{it} + \xi_{it} > 0) \qquad \forall i = 1, ..., N; t = 1, ..., T-1$$

where X_{it} is a vector of observed demographic and family structure variable U_{it} captures the effects of unobserved factors, and β and γ are parameters. There are two latent sources for the unobserved term U_{it}:

$$U_{it} = V_i + \rho\varepsilon_{it-1}$$

where V_i is an individual-specific component, which captures unobserved time invariant factors possibly correlated with the time-varying X_{it}'s such as skill level or motivation; ε_{it} is a serially correlated error term, which captures factors such as transitory wage movements.

In order to provide comparison of the estimators developed in this paper and by Hyslop (1999), we also use the data related to waves 12-19 of the Michigan Panel Survey of Income Dynamics from the calendar years 1979-85 to study married women's employment decisions. The seven-year sample consists of women aged 18-60 in 1980, continuously married, and the husband is a labor force participant in each of the sample years. A woman is defined to be a labor market participant if she works for money any time in the sample year.[27] We obtain a sample having 1752 married women.[28]

As the identification of the models hinges on assumptions in Section 7.1.3, a careful discussion of them in this labor force application is necessary, while we realize that testing these assumptions is not feasible as discussed before. Assumption 7.1.1 is a model specification

[27] A standard definition of a participant is that an individual reports both positive annual hours worked and annual earnings. Hyslop (1995) provided a description of the extent of aggregation bias which results from ignoring intra-year labor force transition.

[28] Hyslop (1995) obtains a sample consisted of 1812 observations. The descriptive statistics of our sample is very close to Hyslop (1995).

and it implies that regardless of whatever is in X_{it}, Y_{it-1}, and U_{it}, enough information has been included so that further lags of participation decision and the explanatory variables including nonlabor income, fertility status, etc, do not matter for explaining the current participation decision Y_{it} directly. Assumption 7.1.3 imposes functional form restrictions on the covariate evolution and the initial joint distribution. Assumption 7.1.4 in the empirical application may be $E\left[I\left(Y_{it}=0\right)|x_{it}, y_{it-1}, u_{it}\right] = F_{\xi_{it}}\left[-\left(x_{it}'\beta + \gamma y_{it-1} + u_{it}\right)\right]$, which is decreasing in u_{it}. Since u_{it} can represent or contain unobserved heterogeneity such as individual ability or motivation, the assumption suggests that the conditional expectation of absence from labor force decreases with ability or motivation. Our choice of G in Assumption 7.1.5 is the mode since the covariate X_{it} contains income variables. In Current Population Survey (CPS), it was found that the mode of misreported income conditional on true income is equal to the true income (see Bound and Krueger (1991) and Chen, Hong, and Tarozzi (2008)). Using the mode condition may relieve concerns on measurement errors. Obviously, this is not the only choice of the functional G. As discussed before, we may use mean or median as well.

We then focus on Assumption 7.1.2. The discussion of the assumption in Section 7.1.3 suggests that it imposes the key restriction that conditional on X_{it} and U_{it}, X_{it+1} is independent of the exogenous shock ξ_{it} and the lagged effects of Y_{it} enter the evolution of X_{it+1} through U_{it}. The regressors of interest in this empirical application are the nonlabor income variables and the fertility variables. There are several scenarios for the exogenous participation shock ξ_{it}. First, if ξ_{it} denotes measurement error, then the conditional independence between ξ_{it} and the future nonlabor income and fertility variables is plausible. Second, if ξ_{it} represents luck in labor markets such as unexpected change of child-care cost or fringe benefit for married women from working, the assumption rules out the immediate effect of the current shock ξ_{it} on the future nonlabor income and fertility variables. This implies that married women do not adjust their nonlabor income and fertility variables to the latest participation shock ξ_{it} but consider all other past period information. If there was a negative shock on participation, married women's nonlabor income and fertility decisions would wait one period to respond to it. Therefore, Assumption 7.1.2 may be plausible in our model of the intertemporal labor force participation behavior of married women. Nevertheless, Assumption 7.1.2 does rule out the possible correlation between the fertility decisions in X_{it+1} and a negative shock on labor force participation ξ_{it} even conditioning on the fertility decisions in the previous period in X_{it}. While the lagged effects of Y_{it} enter the evolution of X_{it+1} indirectly here, our identification strategy still applies with $f_{X_{it+1}|Y_{it},X_{it},Y_{it-1},X_{it-1},U_{it}} = f_{X_{it+1}|X_{it},Y_{it-1},U_{it}}$ in Assumption 7.1.2 if Y_{it-1} has direct influence on X_{it+1}. This alternative specification implies that the labor force participation in period $t-1$ affect married women's future nonlabor income and fertility decisions.

We then apply the sieve MLE method introduced in Section 7.1.4 & 7.1.5 and maintain a single-index form and a mode condition. The estimation results for the various models of labor force participation are presented in Table 7.5 which includes estimates from static probit models with random effect (column 1), a maximum simulated likelihood (MSL) es-

timator[29] (column 2), and the sieve MLE estimator (column 3) for dynamic models. All specifications include unrestricted time effects, a quadratic in age, race, years of education, permanent and transitory nonlabor income y_{mp} & y_{mt}, current realizations of the number of children aged 0-2, 3-5, and 6-17, and lagged realizations of the number of children aged 0-2.[30] While the first two estimators are estimated using full seven years of data, the last one is estimated over three periods of data. In addition, the last estimator is for the dynamic model without an initial conditions specification. The static probit model is estimated by MSL with 200 replications. It allows for individual-specific random effects but ignores possible dynamic effects of the past employment and potential correlation between the unobserved heterogeneity and the regressors. The estimation results of coefficients and APEs indicate that permanent nonlabor income has a significantly negative effect, transitory income reduces the contemporaneous participation, and preschool children have substantially negative effect. In addition, the variance of unobserved heterogeneity is 0.786. We now turn to dynamic specifications. The specifications in the MSL estimator contain random effects, a stationary AR(1) error component, and first-order state dependence (SD(1)). The estimated coefficients and APEs share a similar pattern. The APE estimates show a large and significant first-order state dependence effect reduces the labor force participation probability by about 0.325. The addition of SD(1) and AR(1) error component greatly reduced the effects of nonlabor income variables (-0.002 & -0.001) and the contemporaneous fertility variables like #Kid3-5$_t$ and #Kid6-17$_t$. But the estimated effects of younger kids in the past and current periods #Kid0-2$_{t-1}$ and #Kid0-2$_t$ have stronger negative effects on the probability of women's participation decisions (-0.036 & -0.112). Including state dependence and serial correlation error component reduce the error variance (0.313) due to unobserved heterogeneity. The estimated AR(1) coefficient ρ is -0.146.[31]

[29] A detailed discussion of MSL estimators can be found in Hyslop (1999). There are more specifications in the paper. Here we only compare the models allowing the three sources of persistence.

[30] The labor earnings of the husband are used as a proxy for nonlabor income. Permanent nonlabor income y_{mp} is estimated by the sample average, and transitory income y_{mt} is measured as deviations from the sample average

[31] A correlated random-effects (CRE) is adopted in Hyslop (1999) to test the exogeneity of fertility with respect to participation decisions. His results show that there is no evidence against the exogeneity of fertility decision in dynamic model specifications.

Table 7.5: Estimates of Married Women's Participation Outcomes

	Static Probit+RE (1)		MSL, RE AR(1)+SD(1) (2)		Semi-parametric Probit (3)	
	Coefficient	APE	Coefficient	APE	Coefficient	APE
y_{t-1}	–	–	1.117	0.325	1.089	0.225
	–	–	(0.528)	(0.015)	(0.077)	(0.014)
y_{mp}	-0.312	-0.070	-0.007	-0.002	-0.221	-0.048
	(0.045)	(0.005)	(0.017)	(0.001)	(0.012)	(0.003)
y_{mt}	-0.106	-0.024	-0.004	-0.001	-0.106	-0.023
	(0.026)	(0.002)	(0.028)	(0.001)	(0.056)	(0.001)
#Kid0-2$_{t-1}$	-0.022	-0.005	-0.117	-0.036	-0.055	-0.012
	(0.010)	(0.001)	(0.013)	(0.002)	(0.048)	(0.001)
#Kid0-2$_t$	-0.330	-0.070	-0.380	-0.112	-0.316	-0.065
	(0.021)	(0.005)	(0.145)	(0.006)	(0.061)	(0.004)
#Kid3-5$_t$	-0.400	-0.086	-0.206	-0.062	-0.137	-0.029
	(0.015)	(0.007)	(0.027)	(0.003)	(0.028)	(0.002)
#Kid6-17$_t$	-0.120	-0.028	-0.056	-0.018	-0.062	-0.014
	(0.011)	(0.002)	(0.037)	(0.001)	(0.011)	(0.001)
Cov. Parameters						
σ_v^2	0.786	–	0.313	–	–	–
	(0.071)	–	(0.323)	–	–	–
ρ	–	–	-0.146	–	–	–
	–	–	(0.140)	–	–	–

Note: Bootstrap (simulation) standard errors are reported in parentheses, using 100 bootstrap replications. The models in the first two columns are estimated using full seven years of data but the last two columns are estimated over three-period data. APEs are reported by taking derivatives or differences of ASF at the sample mean of (x_t, y_{t-1}).

The results also show first-order state dependence has a significant positive effect on the probability of participation (0.225). There exists a strong dependence between married women's current labor force participation and past labor force participation, and relaxing the initial conditions assumption increase the negative effects of nonlabor income variables and their significance in the dynamic models. Permanent income and transitory income both reduce the probability of participation but the effect of permanent nonlabor income has substantially greater magnitude.

The fertility variables in the estimation are generally similar to those in column (1) and (2) but with smaller magnitude. That is: each of them has a significantly negative effect on married women's current labor force participation status, and younger children have stronger effect than older. In our semi-parametric probit estimator, the unobserved heterogeneity and the AR(1) component have been mixed into the unobserved covariate U_{it}. They are not identified so there are not any estimation results.

In comparison to the results across specifications allowing for CRE, AR(1), and SD(1), using unspecified CRE and avoiding initial conditions have significant effect on the estimation. The APE estimates find a larger significant negative effects on nonlabor income

variables (-0.048 and -0.002 v.s. -0.023 and -0.001, respectively) and negative effects of children age 0-2 in the current period and past period which decreases by 42% (from -0.112 to -0.065) and decline by 66% (from -0.036 to -0.012) respectively.

7.1.7 Identification in the Discrete Case

We will show how to utilize the identification techniques in Section 2 for the discrete case. The discrete case refers to that the variables X_{it} and U_{it} is discrete:

$$X_{it} \in \mathcal{X}_t \equiv \{1, 2, ..., J_1\} \text{ and } U_{it} \in \mathcal{U} \equiv \{1, 2, ..., J_2\}.$$

In this finite dimensional discrete example, linear integral operators are matrices, which might be useful to give some intuition about how the identification is achieved. For simplicity, assume that $J_1 = J_2 = J$. Based on Eq. (7.13) which is the consequence of Assumption 7.1.1 and 7.1.2, the key equation of the discrete case is:

$$f_{X_{it+1},Y_{it},X_{it},Y_{it-1},X_{it-1}} = \sum_{U_{it}=1}^{J} f_{X_{it+1}|X_{it},U_{it}} f_{Y_{it}|X_{it},Y_{it-1},U_{it}} f_{X_{it},Y_{it-1},X_{it-1},U_{it}}. \tag{7.25}$$

Given $(y_{it}, x_{it}, y_{it-1})$, define J-by-J matrices

$$\begin{aligned}
L_{X_{it+1},y_{it},x_{it},y_{it-1},X_{it-1}} &= \left[f_{X_{it+1},Y_{it},X_{it},Y_{it-1},X_{it-1}}(u, y_{it}, x_{it}, y_{it-1}, x)\right]_{u,x} \\
L_{X_{it+1},x_{it},y_{it-1},X_{it-1}} &= \left[f_{X_{it+1},X_{it},Y_{it-1},X_{it-1}}(u, x_{it}, y_{it-1}, x)\right]_{u,x} \\
L_{X_{it+1}|x_{it},U_{it}} &= \left[f_{X_{it+1}|X_{it},U_{it}}(x|x_{it}, u)\right]_{x,u} \\
L_{x_{it},y_{it-1},X_{it-1},U_{it}} &= \left[f_{X_{it},Y_{it-1},X_{it-1},U_{it}}(x_{it}, y_{it-1}, x, u)\right]_{u,x}
\end{aligned}$$

and a J-by-J diagonal matrix

$$D_{y_{it}|x_{it},y_{it-1},U_{it}} = \begin{bmatrix} f_{Y_{it}|X_{it},Y_{it-1},U_{it}}(y_{it}|x_{it}, y_{it-1}, 1) & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & f_{Y_{it}|X_{it},Y_{it-1},U_{it}}(y_{it}|x_{it}, y_{it-1}, J) \end{bmatrix}.$$

Using these matrixes, Eq. (7.25) can be expressed into a matrix notation as

$$L_{X_{it+1},y_{it},x_{it},y_{it-1},X_{it-1}} = L_{X_{it+1}|x_{it},U_{it}} D_{y_{it}|x_{it},y_{it-1},U_{it}} L_{x_{it},y_{it-1},X_{it-1},U_{it}}. \tag{7.26}$$

Integrating out Y_{it} in Eq. (7.25) leads to

$$f_{X_{it+1},X_{it},Y_{it-1},X_{it-1}} = \sum_{u_{it}=1}^{J} f_{X_{it+1}|X_{it},U_{it}} f_{X_{it},Y_{it-1},X_{it-1},U_{it}}. \tag{7.27}$$

which is equivalent to

$$L_{X_{it+1},x_{it},y_{it-1},X_{it-1}} = L_{X_{it+1}|x_{it},U_{it}} L_{x_{it},y_{it-1},X_{it-1},U_{it}}. \tag{7.28}$$

Assumption 7.1.3 guarantees that the above matrix $L_{X_{it+1},x_{it},y_{it-1},X_{it-1}}$ is invertible. It follows that

$$L_{X_{it+1},y_{it},x_{it},y_{it-1},X_{it-1}} L^{-1}_{X_{it+1},x_{it},y_{it-1},X_{it-1}} = L_{X_{it+1}|x_{it},U_{it}} D_{y_{it}|x_{it},y_{it-1},U_{it}} L^{-1}_{X_{it+1}|x_{it},U_{it}}.$$

The observed matrix on the LHS has a matrix factorization, the product of a diagonal matrix with a matrix of eigenvectors. Uniqueness of the factorization requires the distinct eigenvalues and normalization of the unobserved covariate U_{it}. Assumption 7.1.4 and 7.1.5 are imposed to make these conditions hold. Since the eigenvalues and eigenvectors in the matrix factorization are $f_{Y_{it}|X_{it},Y_{it-1},U_{it}}$ and $f_{X_{it+1}|X_{it},U_{it}}$ respectively, the identification of the model is reached. By Eq. (7.26), the initial joint distribution $f_{X_{it},Y_{it-1},X_{it-1},U_{it}}$ is also identified.

7.2 Misclassification in Treatment Effect Models

7.2.1 Treatment Effect Models

This and the next few sections provide a brief introduction of causal Inference in a treatment effect framework.[32] We are interested in estimating the effect of an intervention on an outcome variable, or vector of outcome variables. Examples of interventions and outcome variables may be

Treatment D	Outcome variable Y
Cancer treatment	Survival time of patients
Job training program	Earnings after training
Change in class size	Students' test scores

Here we only consider a 0-1 binary D. A simple regression model is

$$Y = \alpha + \beta D + \varepsilon$$

This OLS estimator of β is

$$\hat{\beta} = \frac{\sum_{i=1}^{N} Y_i D_i}{\sum_{i=1}^{N} D_i} - \frac{\sum_{i=1}^{N} (1 - D_i) Y_i}{\sum_{i=1}^{N} (1 - D_i)}$$

There are various problems with this approach. First, D and ε may be correlated. Second, the treatment effects may be heterogeneous, i.e., β need not be the same for every member of the population. Third, the OLS estimator is the difference between the average outcome in the treatment and the control group. To analyze this, we consider the so-called Holland-Rubin causal model.

Consider a population with members $i = 1, \ldots, N$. For each member of the population

[32] This part is based on, my thesis advisor, Geert Ridder's lecture notes. All errors are still mine.

we consider $\{Y_i, X_i, D_i\}, i = 1, \ldots, N$ with

$$\begin{aligned} Y &= \text{Outcome variable} \\ D &= \text{Treatment indicator} \\ X &= \text{Vector of other variables} \end{aligned} \tag{7.29}$$

Potential or latent outcomes

$$\begin{aligned} Y_i(0) &= \text{Outcome if } i \text{ is not treated} \\ Y_i(1) &= \text{Outcome if } i \text{ is treated} \end{aligned}$$

$Y(1)$ is called the treated outcome and $Y(0)$ the control outcome. And the treatment effect for individual i is defined as

$$Y_i(1) - Y_i(0)$$

However, for each i we cannot observed *both* $Y_i(0)$ and $Y_i(1)$, but only

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

In other words: For each individual i, either $Y_i(0)$ (if $D_i = 1$) or $Y_i(1)$ (if $D_i = 0$) is missing. Furthermore, we assume that X_i is always observed and unaffected by treatment. We refer to such variables as covariates, regressors, or independent variables. We also assume treatment of i only affects i and not $j \neq i$.

Since the individual treatment effect $Y_i(1) - Y_i(0)$ cannot be estimated, the best we can hope for is to recover the marginal distributions of $Y(1)$ and $Y(0)$ or quantities that can be found from these marginal distributions, i.e. parameters defined on these marginal distributions. We can also derive bounds on the cdf and quantiles of the distribution of individual treatment effects. There are two popular parameters of interest:

- Average Treatment Effect (ATE):

$$E[Y(1) - Y(0)] = \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0)) = \frac{1}{N} \sum_{i=1}^{N} Y_i(1) - \frac{1}{N} \sum_{i=1}^{N} Y_i(0)$$

- Average Treatment Effect on the Treated (ATET)

$$\begin{aligned} E[Y(1) - Y(0)|D = 1] &= \frac{1}{\sum_{i=1}^{N} D_i} \sum_{i=1}^{N} D_i [Y_i(1) - Y_i(0)] \\ &= \frac{\sum_{i=1}^{N} D_i Y_i(1)}{\sum_{i=1}^{N} D_i} - \frac{\sum_{i=1}^{N} D_i Y_i(0)}{\sum_{i=1}^{N} D_i} \\ &= \frac{\sum_{i=1}^{N} D_i Y_i}{\sum_{i=1}^{N} D_i} - \frac{\sum_{i=1}^{N} D_i Y_i(0)}{\sum_{i=1}^{N} D_i} \end{aligned} \tag{7.30}$$

ATE and ATET are the most common parameters. Note that the ATE is the only feature

of the distribution of $Y(1) - Y(0)$ that can be recovered from the marginal distributions of $Y(0)$ and $Y(1)$.

The observed treatment effect, i.e., the simple OLS, can be decomposed as follows:

$$\begin{aligned} & E(Y|D=1) - E(Y|D=0) \\ = \; & E(Y(1)|D=1) - E(Y(0)|D=0) \\ = \; & E(Y(1) - Y(0)|D=1) + E(Y(0)|D=1) - E(Y(0)|D=0) \end{aligned} \tag{7.31}$$

where $E(Y(1) - Y(0)|D = 1)$ is the ATET and $E(Y(0)|D = 1) - E(Y(0)|D = 0)$ is the so-called *selection effect*. The key to estimation of ATE or ATET is to find a way to deal with the selection effect, $E(Y(0)|D = 1) - E(Y(0)|D = 0)$.

There are a few existing solutions, or assumptions, that are known to work

1. Direct randomization,
2. Conditional randomization: Unconfounded assignment or conditional independence,
3. Indirect randomization: Instrumental variables,
4. Local randomization: Regression discontinuity,
5. Second-order randomization: Difference-in-difference.

7.2.2 Direct Randomization

Randomized assignment of treatment means

$$\{Y(0), Y(1)\} \perp D$$

This implies mean independence

$$E[Y(0)|D = 1] = E[Y(0)]$$

$$E[Y(1)|D = 1] = E[Y(1)]$$

Hence

$$E[Y(0)|D = 1] - E[Y(0)|D = 0] = 0$$

i.e. the selection effect is 0. In addition,

$$E[Y(1) - Y(0)|D = 1] = E[Y(1) - Y(0)]$$

That means ATET=ATE.

7.2.3 Conditional Randomization: Unconfounded Assignment

Unconfounded assignment means that

$$\{Y(1), Y(0)\} \perp D|X.$$

This assumption implies that X contains all variables that affect both the selection for treatment and the outcomes. With unconfounded assignment $\{Y(1), Y(0)\} \perp D|X$, we have

$$E(Y(0)|X, D = 1) - E(Y(0)|X, D = 0) = 0$$

and

$$E(Y(1) - Y(0)|X, D = 1) = E(Y(1) - Y(0)|X)$$

Hence

$$E(Y|X, D = 1) - E(Y|X, D = 0) = E(Y(1) - Y(0)|X)$$

In other words, observed treatment effect given X is equal to the ATE and ATET given X. By the law of iterated expectations, the unconditional treatment effects are

$$E[Y(1) - Y(0)] = E_x \left[E(Y(1) - Y(0)|X)\right]$$

and

$$E[Y(1) - Y(0)|D = 1] = E_x \left[E(Y(1) - Y(0)|X) \,|D = 1\right]$$

Note that in general ATE $\neq$ ATET.

Another issue is that non-parametric regressions $E(Y|X, D = 1)$, $E(Y|X, D = 0)$ suffer from curse of dimensionality. The question is whether we can summarize X is a lower dimension. Define

$$p(X) = \Pr(D = 1|X)$$

This is the *probability of selection* or *propensity score*. Rosenbaum and Rubin (1983) proved two results

- Balancing score property
- Sufficiency of the propensity score for unconfoundedness

Balancing score property or sufficiency of the propensity score for D

$$D \perp X|p(X)$$

Proof: Because

$$X = x \Rightarrow p(X) = p(x)$$

we have

$$\Pr(D = 1|X = x, p(X) = p(x)) =$$

$$= \Pr(D = 1|X = x) = p(x)$$

By the law of iterated expectations

$$\Pr(D = 1|p(X) = p(x)) =$$

$$= E_x \left[\Pr(D = 1|X, p(X) = p(x))|p(X) = p(x)\right] =$$

$$= E_x \left[p(X)|p(X) = p(x)\right] = p(x)$$

Hence

$$\Pr(D = 1|X = x, p(X) = p(x)) =$$

$$= \Pr(D = 1|p(X) = p(x))$$

so that $D \perp X|p(X)$ □.

Implication: In regression of D on $X, p(X)$, the coefficients of X are 0 (or not significantly different from 0).

Unconfounded assignment given the propensity score

$$Y(0), Y(1) \perp D|X \Rightarrow Y(0), Y(1) \perp D|p(X)$$

Proof: We have

$$\Pr(D = 1|Y(0), Y(1), p(X) = p(x)) =$$

$$= E_x \left[\Pr(D = 1|Y(0), Y(1), p(X) = p(x), X)|Y(0), Y(1), p(X) = p(x)\right] =$$

$$= E_x \left[\Pr(D = 1|Y(0), Y(1), X)|Y(0), Y(1), p(X) = p(x)\right] =$$

$$= E_x \left[p(X)|Y(0), Y(1), p(X) = p(x)\right] = p(x) =$$

$$= \Pr(D = 1|p(X) = p(x))$$

□.

By this result the observed treatment effect given $p(X)$ is

$$E(Y|p(X), D = 1) - E(Y|p(X), D = 0) =$$

$$= E(Y(1) - Y(0)|p(X), D = 1) +$$

$$+E(Y(0)|p(X), D = 1) - E(Y(0)|p(X), D = 0) =$$

$$= E(Y(1) - Y(0)|p(X))$$

Define $P = p(X)$, then

$$\text{ATE} = E_p \left[E(Y|P, D = 1) - E(Y|P, D = 0)\right]$$

and

$$\text{ATET} = E_p\left[E(Y|P, D = 1) - E(Y|P, D = 0)|D = 1\right]$$

$E(Y|P, D = 1), E(Y|P, D = 0)$ are non-parametric regressions on one variable P instead of a vector X.

Based on these observations, we can propose a simple non-parametric estimator: propensity score matching. We subdivide interval $[0, 1]$ into $0 = p_0 < p_1 < p_2 < \ldots < p_{K-1} < p_K = 1$ and compute

$$m_{k1} = \frac{\sum_{i=1}^N D_i I(p_{k-1} \leq P_i < p_k) Y_i}{\sum_{i=1}^N D_i I(p_{k-1} \leq P_i < p_k)}$$

$$m_{k0} = \frac{\sum_{i=1}^N (1 - D_i) I(p_{k-1} \leq P_i < p_k) Y_i}{\sum_{i=1}^N (1 - D_i) I(p_{k-1} \leq P_i < p_k)}$$

Then

$$\text{ATE} = \sum_{k=1}^K \frac{\sum_{i=1}^N I(p_{k-1} \leq P_i < p_k)}{N} (m_{k1} - m_{k0})$$

$$\text{ATET} = \frac{\sum_{i=1}^N Y_i D_i}{\sum_{i=1}^N D_i} - \sum_{k=1}^K \frac{\sum_{i=1}^N I(p_{k-1} \leq P_i < p_k)}{N} m_{k0}$$

Estimation as Missing Data Models

Key problem for causal inference is the fact that the counterfactual outcome is missing: for the treated we do not observe the non-treated outcome Y_0 and for the non-treated we do not observed the treated outcome Y_1. Analogy with the missing data problem where we observe $\{D, D \times Y, X\}$, D is the observation indicator, and 0 is the label for missing Y. If $Y \equiv Y_1$, then we do not observe Y_1 for the non-treated $D = 0$. Hence, the causal inference problem is equivalent to two missing data problems with $Y \equiv Y_1$ and $Y \equiv Y_2$, respectively. The assumption of unconfounded assignment is equivalent to the assumption

$$Y \perp D|X$$

i.e. the assumption that Y is Missing-at-Random (MAR). The parameter of interest in the missing data problem is $E(Y)$, and the question becomes: what is the best estimator of $E(Y)$ if we observe $\{D, D \times Y, X\}$.

If we observe Y directly, i.e. if we have a random sample Y_i for $i = 1, 2, \ldots, N$, then the sample mean

$$\frac{1}{N} \sum_{i=1}^N Y_i$$

is the (semi-)parametrically efficient estimator of $E(Y)$ (if $E(Y^2) < \infty$, i.e. it has the smallest variance. Same argument as sufficiency of propensity score for treatment assignment gives

$$D \perp Y|X \Rightarrow D \perp Y|p(X)$$

with $p(X) = \Pr(D = 1|X)$.

Using this, and the law of iterated expectations, we have the identity for $E(Y)$

$$E(Y) = E_{p(X)}[E_Y(Y|p(X))] = \tag{7.32}$$

$$= E_{p(X)}[E_Y(Y|D = 1, p(X))]$$

An alternative identity is

$$E\left(\frac{D \times Y}{p(X)}\right) = E_{D,p(X)}\left[E_Y\left(\frac{D \times Y}{p(X)}\Big|D, p(X)\right)\right] = \tag{7.33}$$

$$= E_{p(X)}\left[\frac{E_Y(Y|p(X))\Pr(D = 1|p(X))}{p(X)}\right] =$$

$$= E_{p(X)}\left[E_Y(Y|p(X)\right] = E(Y)$$

because by $D \perp X|p(X)$, $\Pr(D = 1|X, p(X)) = \Pr(D = 1|X) = p(X)$.
The two identities correspond to two methods to estimate $E(Y)$

- Regress Y in the sample, i.e. given D=1, on $p(X)$. Average the predicted value $E(Y|D = 1, p(X))$ over $p(X)$. This is the *imputation* estimator with imputed missing potential outcomes

$$\hat{Y}_i(1) = \frac{\hat{E}(DY|X_i)}{\hat{p}(X_i)},$$

$$\hat{Y}_i(0) = \frac{\hat{E}((1 - D)Y|X_i)}{1 - \hat{p}(X_i)}.$$

  Hahn (1998) shows that this imputation leads to a semi-parametrically efficient estimator of treatment effects.

- Weight the observed Y by $\frac{1}{p(X)}$, i.e. the inverse probability of observation, and take the average. This the *weighting* estimator.

  For a sample $\{D_i, D_i.Y_i, X_i\}, i = 1, \ldots, N$, the estimator is

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\frac{D_i \times Y_i}{\hat{p}(X_i)}$$

  Note that the weight must be estimated. Hirano et al. (2003) show that this estimator leads to an semiparametric efficient estimator of treatment effects, i.e.,

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i \times Y_i}{\hat{p}(X_i)} - \frac{(1 - D_i) \times Y_i}{1 - \hat{p}(X_i)}\right)$$

7.2.4 Indirect Randomization: Instrumental Variables

A simple linear regression model of relation between treatment and outcome is

$$Y = \alpha + \beta D + u.$$

This corresponds to constant treatment effect model with potential outcomes

$$\begin{aligned} Y_0 &= \alpha + u \\ Y_1 &= \alpha + \beta + u \end{aligned}$$

The constant ATE=ATET is β if $D \perp u$ or $E(u|D = 0) = E(u|D = 1)$ Then the OLS estimator of β

$$\hat{\beta} = \frac{\sum_{i=1}^N D_i Y_i}{\sum_{i=1}^N D_i} - \frac{\sum_{i=1}^N (1 - D_i) Y_i}{\sum_{i=1}^N (1 - D_i)}$$

is unbiased.

When the treatment is endogenous or a self-selected choice, we have

$$E(u|D = 0) \neq E(u|D = 1)$$

i.e., the regressor D is endogenous. One solution is to introduce an instrument. We consider an instrument Z, which is a 0-1 variable with

- $E(u|Z = 1) = E(u|Z = 0)$
- $\Pr(D = 1|Z = 0) \neq \Pr(D = 1|Z = 1)$

Then the IV or 2SLS estimator of β is

$$\hat{\beta}_{IV} = \frac{\frac{\sum_{i=1}^N Y_i Z_i}{\sum_{i=1}^N Z_i} - \frac{\sum_{i=1}^N Y_i (1-Z_i)}{\sum_{i=1}^N (1-Z_i)}}{\frac{\sum_{i=1}^N D_i Z_i}{\sum_{i=1}^N Z_i} - \frac{\sum_{i=1}^N D_i (1-Z_i)}{\sum_{i=1}^N (1-Z_i)}}$$

This converges in probability to

$$plim\, \hat{\beta}_{IV} = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{\Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)} =$$

$$= \frac{\beta E(D|Z = 1) - \beta E(D|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)} = \beta$$

One application is Angrist (1990), which investigates the effect of participation in Vietnam war on lifetime earnings . The treatment is

$$\begin{aligned} D &= 1 \quad \text{if in Vietnam war} \\ &= 0 \quad \text{if not} \end{aligned}$$

The outcome variable is

$$Y = \text{lifetime earnings}$$

The observed effect is

$$E(Y|D=1) - E(Y|D=0)$$

But this may not be equal to ATE or ATET because

$$\begin{aligned} & E(Y|D=1) - E(Y|D=0) \\ = \; & E(Y_1 - Y_0|D=1) + E(Y_0|D=1) - E(Y_0|D=0) \end{aligned} \tag{7.34}$$

The first term on the right-hand side is ATET, the second term is the selection effect. For the constant treatment effect model this is equal to

$$\beta + E(u|D=1) - E(u|D=0)$$

The instrument Z is the randomized treatment assignment

- Conscription with all men born in 1950-1953 potential conscripts.
- Not all men in birth cohort needed.
- For reasons of equity all potential conscripts should have same chance of being drafted: Draft lottery.
- Example: 1971 lottery for men born in 1951.
  - In December 1970 random numbers 1-365 assigned to days in 1971, e.g. January 1 is 33, January 2 is 288, etc.
  - Based on need of army, all men born on a date with a random number below 125 were drafted.

The instrument is defined as follows:

$$\begin{aligned} Z &= 1 \quad \text{if draft eligible} \\ &= 0 \quad \text{if not} \end{aligned}$$

Note that

$$\begin{aligned} Z = 1, D = 0 \quad &\rightarrow \quad \text{draft avoiders} \\ Z = 0, D = 1 \quad &\rightarrow \quad \text{volunteers} \end{aligned}$$

In this application, it is obvious that

$$Z \perp \{Y_0, Y_1\}$$

That data also show

$$\Pr(D=1|Z=1) \neq \Pr(D=1|Z=0)$$

Table 7.6: Endogenous treatment with randomized assignment as IV

	Randomized assignment $Z = z$		
	$z = 0$	$z = 1$	Type
Endogenous	$D(z) = 0$	$D(z) = 0$	Never takers
treatment	$D(z) = 0$	$D(z) = 1$	Compliers
or	$D(z) = 1$	$D(z) = 0$	Deniers
choice $D(Z)$	$D(z) = 1$	$D(z) = 1$	Always takers

Hence we use the IV estimator with the 0-1 Z as instrument.

Until now we have been considering constant treatment effect, what would the IV estimator estimate if the treatment effects are heterogeneous? We consider the case where treatment Z is randomly assigned and individuals make the choice of whether to take the treatment, i.e. there is non-compliance. We define Y_0, Y_1 as potential outcomes , $D_0 = D(z = 0), D_1 = D(z = 1)$ as potential treatments for $Z = z$. Observed treatment or choice is

$$D = ZD_1 + (1 - Z)D_0$$

In general, Angrist et al. (1996) suggest that there are four types in the population as shown in Table 7.6

We assume

$$Z \perp \{Y_0, Y_1, D_0, D_1\}$$

We have

$$Y = DY_1 + (1 - D)Y_0 =$$

$$= (ZD_1 + (1 - Z)D_0)Y_0 + (1 - ZD_1 - (1 - Z)D_0)Y_0$$

so that

$$\begin{aligned} Y|Z = 1 &\overset{d}{=} Y_0 + D_1(Y_1 - Y_0)|Z = 1 \\ Y|Z = 0 &\overset{d}{=} Y_0 + D_0(Y_1 - Y_0)|Z = 0 \end{aligned}$$

Hence, the (reduced form) effect of Z on Y is

$$E(Y|Z = 1) - E(Y|Z = 0) = E\left[(D_1 - D_0)(Y_1 - Y_0)\right] =$$

$$= E(Y_1 - Y_0|D_1 - D_0 = 1)\Pr(D_1 - D_0 = 1) -$$

$$-E(Y_1 - Y_0|D_1 - D_0 = -1)\Pr(D_1 - D_0 = -1)$$

Monotonicity

$$D_1 \geq D_0 \qquad \text{or} \qquad D_1 \leq D_0$$

Hence for all members of population $i = 1, \ldots, I$

$$D_{1i} \geq D_{0i} \qquad \text{or} \qquad D_{1i} \leq D_{0i}$$

If Z is treatment assignment and D treatment choice, then this is equivalent to

- i chooses non-treatment when assigned treatment, i.e. $D_{1i} = 0 \Rightarrow i$ chooses non-treatment when not assigned treatment, i.e. $D_{0i} = 0$
- i chooses treatment when not assigned treatment, i.e. $D_{0i} = 1 \Rightarrow i$ chooses treatment when assigned treatment, i.e. $D_{1i} = 0$

Because this involves counterfactual treatments, monotonicity not testable. Assume $D_1 \geq D_0$, i.e. there is no “Deniers". Then

$$E(Y|Z=1) - E(Y|Z=0) = E\left[(D_1 - D_0)(Y_1 - Y_0)\right] =$$

$$= E(Y_1 - Y_0|D_1 - D_0 = 1)\Pr(D_1 - D_0 = 1)$$

Hence

$$E(Y_1 - Y_0|D_1 - D_0 = 1) = \frac{E(Y|Z=1) - E(Y|Z=0)}{\Pr(D_1 - D_0 = 1)}$$

Because $D_1 - D_0$ is 0-1

$$\Pr(D_1 - D_0 = 1) = E(D_1 - D_0) =$$

$$= Pr(D_1 = 1) - Pr(D_0 = 1)$$

Hence

$$E(Y_1 - Y_0|D_1 = 1, D_0 = 0) = \frac{E(Y|Z=1) - E(Y|Z=0)}{\Pr(D_1 = 1) - \Pr(D_0 = 1)}$$

The left-hand side is the so-called Local Average Treatment Effect (LATE).

Consider potential outcomes

$$\begin{aligned} Y_0 &= \alpha + u \\ Y_1 &= \alpha + \beta + \eta + u \end{aligned}$$

with $E(\eta) = 0$. Individual treatment effect is

$$Y_1 - Y_0 = \beta + \eta$$

Assume

$$\eta \perp \{D_0, D_1\} \tag{7.35}$$

i.e. if D_0, D_1 are treatment choices (for $Z = 0, 1$), then these choices are independent of the effect heterogeneity, i.e. units do not choose on their individual effect.

Under this assumption

$$E(Y|Z=1) - E(Y|Z=0) = E\left[(D_1 - D_0)(Y_1 - Y_0)\right] =$$

$$\beta E(D_1 - D_0) + E\left[(D_1 - D_0)\eta\right] =$$

$$\beta(\Pr(D_1 = 1) - \Pr(D_0 = 1))$$

Because D is a function of D_0, D_1, Z also

$$\beta = E(Y_1 - Y_0) = E(Y_1 - Y_0|D = 1)$$

Hence LATE=ATE=ATET. This is not true if the assumption in equation (7.35) does not hold!

7.2.5 IV and Marginal Treatment Effects

Our latent variable model for potential outcomes and treatment assignment is specified as follows:

$$\begin{aligned} Y_0 &= \mu_0 + U_0 \\ Y_1 &= \mu_1 + U_1 \\ D^* &= \gamma_0 + \gamma_1 Z - U_D \\ D &= I(D^* \geq 0) \end{aligned}$$

The random errors U_0, U_1, U_D have a joint distribution with U_0, U_D and U_1, U_D correlated. Because either Y_0 or Y_1 are observed, we have no information on the joint distribution of U_0, U_1. The individual treatment effect equals

$$Y_0 - Y_1 = (\mu_0 - \mu_1) + (U_1 - U_0)$$

with $U_1 - U_0$ being the effect heterogeneity. The ATE is

$$E(Y_1 - Y_0) = \mu_1 - \mu_0$$

And the ATET is

$$E(Y_1 - Y_0|D = 1) = \mu_1 - \mu_0 + E(U_1 - U_0|U_D \leq \gamma_0 + \gamma_1 Z)$$

Note that

$$D(z) = I(U_D \leq \gamma_0 + \gamma_1 z)$$

i.e. assignment for different levels of z. We change the notation from D_z to $D(z)$. If $z' > z$ and $\gamma_1 > 0$, then LATE is

$$E(Y_1 - Y_0|D(z) = 0, D(z') = 1) = \mu_1 - \mu_0 +$$

$$+E(U_1 - U_0|\gamma_0 + \gamma_1 z < U_D \leq \gamma_0 + \gamma_1 z')$$

Note that $\{D(z) = 1, D(z') = 0\} \Leftrightarrow \{U_D \leq \gamma_0 + \gamma_1 z, U_D > \gamma_0 + \gamma_1 z'\}$. This latter event is an empty set, which implies that in latent variable model monotonicity always holds.

The marginal treatment effect (MTE) is defined as

$$E(Y_1 - Y_0|U_D = u) = \mu_1 - \mu_0 + E(U_1 - U_0|U_D = u).$$

Below we should how it is related with ATE, ATET, and LATE. Note that

$$f(v_2|v_1 \leq V_1 \leq v_1 + \varepsilon) = \frac{\int_{v_1}^{v_1+\varepsilon} f(s, v_2)\mathrm{d}s}{\int_{v_1}^{v_1+\varepsilon} f(s)\mathrm{d}s}$$

For $\varepsilon \downarrow 0$

$$\lim_{\varepsilon \downarrow 0} f(v_2|v_1 \leq V_1 \leq v_1 + \varepsilon) = f(v_2|v_1)$$

Hence for LATE

$$\lim_{z' - z \downarrow 0} E(Y_1 - Y_0|D(z) = 0, D(z') = 1) =$$

$$= \mu_1 - \mu_0 + E(U_1 - U_0|U_D = \gamma_0 + \gamma_1 z)$$

This is the MTE for $U_D = \gamma_0 + \gamma_1 z$, which is the ATE for units that are at the margin between treatment and non-treatment. We may integrate the MTE to obtain LATE.

$$E(Y_1 - Y_0|D(z) = 0, D(z') = 1) = \mu_1 - \mu_0 +$$

$$+ \int_{\gamma_0 + \gamma_1 z}^{\gamma_0 + \gamma_1 z'} E(U_1 - U_0|U_D = u) f_D(u|\gamma_0 + \gamma_1 z \leq U_D \leq \gamma_0 + \gamma_1 z')\mathrm{d}u$$

For ATET

$$E(Y_1 - Y_0|D(z) = 1) = \mu_1 - \mu_0 +$$

$$+ \int_{-\infty}^{\gamma_0+\gamma_1 z} E(U_1 - U_0|U_D = u) f_D(u|U_D \leq \gamma_0 + \gamma_1 z) \mathrm{d}u$$

For ATE

$$E(Y_1 - Y_0) = \mu_1 - \mu_0 +$$

$$+ \int_{-\infty}^{\infty} E(U_1 - U_0|U_D = u) f_D(u) \mathrm{d}u = \mu_1 - \mu_0$$

because $E(U_1 - U_0) = 0$. If $E(U_1 - U_0|U_D = u) = 0$, i.e. U_D is mean independent of $U_1 - U_0$, then ATE=ATET=LATE=MTE.

Furthermore, because

$$P(z) = \Pr(D = 1|Z = z) = F_D(\gamma_0 + \gamma_1 z)$$

we may have

$$\gamma_0 + \gamma_1 z = F_D^{-1}(p(z))$$

All expressions for treatment effects above are functions of $\gamma_0 + \gamma_1 z$ and hence of $p(z)$. This property of the latent variable selection model is called *index sufficiency*, which is comparable with the sufficiency of the propensity score for unconfounded treatment assignment. However, index sufficiency is a restriction. There are models in which index sufficiency does not hold.

Under index sufficiency, we may estimate the treatment effects based on the estimation of MTE. We consider ATE

$$E(Y_1 - Y_0) =$$

$$\int_{-\infty}^{\infty} \left(\mu_1 - \mu_0 + E(U_1 - U_0|U_D = u)\right) f_D(u) \mathrm{d}u =$$

$$= \int_{-\infty}^{\infty} \mathrm{MTE}(u) f_D(u) \mathrm{d}u = \int_0^1 \mathrm{MTE}(p) \mathrm{d}p$$

ATET given $Z = z$

$$E(Y_1 - Y_0|D(z) = 1, Z = z) =$$

$$\int_{-\infty}^{\gamma_0+\gamma_1 z} \left(\mu_1 - \mu_0 + E(U_1 - U_0|U_D = u)\right).$$

$$.f_D(u|U_D \leq \gamma_0 + \gamma_1 z) \mathrm{d}u = \int_0^{p(z)} \mathrm{MTE}(p) \frac{1}{p(z)} \mathrm{d}p$$

so that ATET

$$E(Y_1 - Y_0|D = 1) = E_Z \left[\int_0^{p(Z)} \mathrm{MTE}(p) \frac{1}{p(Z)} \mathrm{d}p \mid D = 1 \right]$$

If the instrument z varies in $z_0 \leq z \leq z_1$, then $p(z_0) \leq p \leq p(z_1)$. We can estimate $E(Y|p(Z) = p)$ on this interval by (one-dimensional) non-parametric regression, and the derivative w.r.t p is MTE(p). For ATE we need MTE(p) for $0 \leq p \leq 1$ and for ATET we need MTE(p) for $0 \leq p \leq p(z_1)$

The key issue is not variation in p, but the extremes of p. Assume z_0, z_1 with $p(z_0) = 0$ and $p(z_1) = 1$. Then

$$\begin{aligned} E(Y|Z = z_0) &= p(z_0)E(Y_1|D = 1, Z = z_0)+ \\ &+(1 - p(z_0))E(Y_0|D = 0, Z = z_0) = \\ &= E(Y_0|D(z_0) = 0, Z = z_0) = E(Y_0|Z = z_0) = E(Y_0) \end{aligned}$$

Analogously, we have $E(Y|Z = z_1) = E(Y_1)$ so that we identify ATE.

If only $p(z_0) = 0$, then

$$E(Y|Z = z_0) = E(Y_0)$$

and because

$$\begin{aligned} E(Y_0) &= E(Y_0|D = 1)\Pr(D = 1)+ \\ &+E(Y_0|D = 0)\Pr(D = 0) \end{aligned}$$

we have

$$E(Y_0|D = 1) = \frac{E(Y_0) - E(Y|D = 0)\Pr(D = 0)}{\Pr(D = 1)}$$

Because $E(Y|D = 1) = E(Y_1|D = 1)$ we identify ATET.

However, what to do if we only have $0 < p(z_0) \leq p \leq p(z_1) < 1$? One solution is to make distributional assumptions

$$\begin{pmatrix} U_0 \\ U_1 \\ U_D \end{pmatrix} \sim N\left(0, \begin{bmatrix} \sigma_0^2 & . & \rho_0\sigma_0 \\ . & \sigma_1^2 & \rho_1\sigma_1 \\ \rho_0\sigma_0 & \rho_1\sigma_1 & 1 \end{bmatrix}\right)$$

Then

$$E(U_1 - U_0|U_D) = (\rho_1\sigma_1 - \rho_0\sigma_0)U_D$$

Hence

$$\text{MTE}(p) = (\rho_1\sigma_1 - \rho_0\sigma_0)\Phi^{-1}(p)$$

The parametric assumption allows us to extrapolate MTE(p) from any small interval to $[0, 1]$. LaLonde (1986) shows that this is not a good idea, which stimulated the nonparametric estimation of treatment effect models, such as Dehejia and Wahba (1999).

Another solution is to estimate bounds on the treatment effects. Suppose

$$Y_L \leq Y_0, Y_1 \leq Y_H$$

then

$$Y_L - Y_H \leq \text{MTE}(p) = E(Y_1 - Y_0|p) \leq Y_H - Y_L$$

and

$$\text{ATE} = \int_0^{p(z_0)} \text{MTE}(p)\text{d}p + \int_{p(z_0)}^{p(z_1)} \text{MTE}(p)\text{d}p + \int_{p(z_1)}^{1} \text{MTE}(p)\text{d}p$$

Hence

$$(Y_L - Y_H)[p(z_0) + (1 - p(z_1))] + \int_{p(z_0)}^{p(z_1)} \text{MTE}(p)\text{d}p \leq \text{ATE} \leq$$

$$\leq (Y_H - Y_L)[p(z_0) + (1 - p(z_1))] + \int_{p(z_0)}^{p(z_1} \text{MTE}(p)\text{d}p$$

7.2.6 Local Randomization: Regression Discontinuity

Identification of treatment effects requires an exogenous change in the treatment decision, which is independent of the outcomes. In this section, we consider the case where the treatment assignment changes discontinuously at $Z = z_0$. The continuity of the outcomes then brings in identification of treatment effects around the discontinuous point. There are two types of regression discontinuity designs as follows:

- Sharp design: $D = I(Z \geq z_0)$
- Fuzzy design: $p(z_0+) \neq p(z_0-)$, where

$$p(z_0+) \equiv \lim_{z \downarrow z_0} \Pr(D = 1|Z = z)$$

$$p(z_0-) \equiv \lim_{z \uparrow z_0} \Pr(D = 1|Z = z)$$

.

For simplicity, we assume that the treatment effect is constant, i.e., $Y_1 - Y_0 = \beta$, which implies

$$Y = Y_0 + \beta D.$$

The key assumption for regression discontinuity design is that

Assumption 7.2.1 *$E(Y_0|Z = z)$ is continuous in z_0.*

We then have

$$\begin{aligned} & E(Y|Z = z_0 + \varepsilon) - E(Y|Z = z_0 - \varepsilon) \\ = \; & E(Y_0|Z = z_0 + \varepsilon) - E(Y_0|Z = z_0 - \varepsilon) + \\ & +\beta \times [E(D|Z = z_0 + \varepsilon) - E(D|Z = z_0 - \varepsilon)] \end{aligned}$$

For $\varepsilon \downarrow 0$, Assumption 7.2.1 implies that

$$\lim_{\varepsilon \downarrow 0}[E(Y_0|Z = z_0 + \varepsilon) - E(Y_0|Z = z_0 - \varepsilon)] = 0.$$

so that

$$\begin{aligned} \mu(z_0+) - \mu(z_0-) &\equiv \lim_{\varepsilon \downarrow 0}[E(Y|Z = z_0 + \varepsilon) - E(Y|Z = z_0 - \varepsilon)] \\ &= \beta \times [p(z_0+) - p(z_0-)] \end{aligned} \tag{7.36}$$

Therefore, the treatment effect is identified as

$$\beta = \frac{\mu(z_0+) - \mu(z_0-)}{p(z_0+) - p(z_0-)}$$

In fact, this is comparable with the Wald estimator with 0-1 instrument.

7.2.7 Second-Order Randomization: Difference-in-Difference

When the data contain more information, the randomization can be imposed on the second-order variation of the data. For example, suppose we observe individuals' behavior for different time periods in panel data, instead of in cross-sectional data as before. For individual $i = 1, 2, ..., N$ and two periods t, we observe

$$\{Y_{it}, D_{it}, X_{it}\} \quad t = 1, 2$$

We assume there are only treatment in period 2, i.e.,

$$D_1 = 0$$

For simplicity, we assume constant treatment effects with potential outcomes satisfying

$$Y_{it} = \alpha_t + \beta D_{it} + \eta_i + u_{it}, \quad t = 1, 2$$

with η_i an individual effect. Notice that the treatment D_{it} may be correlated with the individual effect η_i, i.e.,

$$\Pr(D_{i2} = 1|\eta_i) \neq \Pr(D_{i2} = 1)$$

More importantly, we assume that

Assumption 7.2.2 *The error term u_{it} is independent of the treatment, i.e.,*

$$\Pr(D_{i2} = 1|u_{i1}, u_{i2}) = \Pr(D_{i2} = 1).$$

To eliminate the individual effect, we take the first-difference to have

$$Y_{i2} - Y_{i1} = \alpha_2 - \alpha_1 + \beta D_{i2} + u_{i2} - u_{i1}$$

The treatment effect β can be estimated by OLS because Assumption 7.2.2 implies that

$$E(u_{i2} - u_{i1}|D_{i2} = 1) = E(u_{i2} - u_{i1}|D_{i2} = 0) = 0$$

The OLS estimator is unbiased for

$$\begin{aligned}\beta &= E(Y_2 - Y_1|D_2 = 1) - E(Y_2 - Y_1|D_2 = 0) \\ &= \{E(Y_2|D_2 = 1) - E(Y_2|D_2 = 0)\} - \{E(Y_1|D_2 = 1) - E(Y_1|D_2 = 0)\}\end{aligned}$$

The first term on the right-hand side is for period 2 and the second for period 1, i.e. the observed treatment effect for period 2 minus the observed treatment effect for period 1 (pre-treatment). The OLS estimator for the change is the difference-in-differences or dif-in-dif estimator. Notice that this estimator does not required panel data. A repeated cross-section is sufficient.

If we consider the potential outcomes, we have

$$\begin{array}{lll} t = 1 & Y_{01} & \\ t = 2 & Y_{02} & Y_{12} \end{array}$$

For convenience, we let $D \equiv D_2$. In two periods, the researcher observes

$$\begin{aligned} Y_2 &= DY_{12} + (1 - D)Y_{02} \\ Y_1 &= Y_{01} \end{aligned}$$

The key assumption here is

Assumption 7.2.3

$$(Y_{02} - Y_{01}) \quad \perp \quad D$$

or

$$E(Y_{02} - Y_{01}|D = 1) = E(Y_{02} - Y_{01}|D = 0).$$

This assumption means that the treatment assignment can be on basis of non-treated outcome level, but not on basis of change in non-treated outcome. Hence

$$Y_2 - Y_1 = Y_{02} - Y_{01} + D(Y_{12} - Y_{02})$$

and

$$\begin{aligned} E(Y_2 - Y_1|D = 1) &= E(Y_{02} - Y_{01}|D = 1) + E(Y_{12} - Y_{02}|D = 1) \\ E(Y_2 - Y_1|D = 0) &= E(Y_{02} - Y_{01}|D = 0) \end{aligned}$$

with the observed difference equal to

$$E(Y_2 - Y_1|D = 1) - E(Y_2 - Y_1|D = 0) = E(Y_{12} - Y_{02}|D = 1)$$

which is the ATET. Note that $E(Y_2 - Y_1|D = 0)$ is the observed average change in non-

treated outcome for the non-treated population. By the assumption this is also the unobserved change in the average outcome in the non-treated state for the treated population.

7.2.8 Misclassification of Treatments

After the introduction of treatment effect models, this section shows how the results for measurement error models are applicable when treatments are mismeasured. Lewbel (2007) considers identification and estimation of the effect of a mismeasured binary regressor in a nonparametric or semiparametric regression, or the conditional average effect of a binary treatment or policy on some outcome where treatment may be misclassified. Let's consider a simplified version of the model without covariates. Define Y as the outcome variable, T^* is the true binary treatment, and V is an exogenous variable. The research observes a mismeasured binary treatment T instead of the true treatment. The key assumption is that the variable V only affects the true treatment probability but not the treatment effect nor misclassification probability. Lewbel (2007) imposed restrictions on the support of V, such as at least three values in the support, and the relationship between V and an explicit function of mislassification probability to avoid directly imposing conditional independence. In fact, the intuition is better captured by a 3-measurement model satisfying

$$f(Y,T,|V) = \sum_{T^* \in \{0,1\}} f(Y|T^*) f(T|T^*) f(T^*|V)$$

This is similar to the mean regression case in Mahajan (2006), where the key relationship can be described as

$$E(Y|T,V) = \frac{1}{f(T|V)} \sum_{T^* \in \{0,1\}} E(Y|T^*) f(T|T^*) f(T^*|V).$$

In fact, Hui and Walter (1980) considers the same theoretical framework with different interpretation, where T^* is the true binary indictor of whether an individual has certain disease, Y and T are two separate diagnostic tests' binary outcome, and V stands for different subpopulations. In this case, the conditional independence assumptions seem very reasonable, i.e.,

$$Y \perp T \perp V \mid T^*.$$

Let $Y, T, V \in \{0,1\}$. They further specify the likelihood function as follows:

$$\begin{aligned}
& f(Y,T,|V) \qquad (7.37)\\
= & \sum_{T^* \in \{0,1\}} f(Y|T^*) f(T|T^*) f(T^*|V) \\
\equiv & [f_{Y|T^*}(1|0)]^Y [1 - f_{Y|T^*}(1|0)]^{1-Y} [f_{T|T^*}(1|0)]^T [1 - f_{T|T^*}(1|0)]^{1-T} [1 - f_{T^*|V}(1|v)] \\
+ & [1 - f_{Y|T^*}(0|1)]^Y [f_{Y|T^*}(0|1)]^{1-Y} [1 - f_{T|T^*}(0|1)]^T [f_{T|T^*}(0|1)]^{1-T} f_{T^*|V}(1|v)
\end{aligned}$$

Notice that $f_{T^*|V}(1|v)$ stands for the probability of a diseased individual in subpopulation $V = v$, $f_{Y|T^*}(1|0)$ is the false positive rate of test Y, and $f_{Y|T^*}(0|1)$ is the false negative rate of test Y. Similarly, $f_{T|T^*}(1|0)$ and $f_{T|T^*}(0|1)$ are the false positive and false negative rates of test T.

Using their notation, we define

$$p_{gij} = f(Y = i, T = j|V = g)$$

with for $g, i, j \in \{0, 1\}$ (their paper uses $g, i, j \in \{1, 2\}$) the true probability of test outcomes i in test Y and j in test T. Let the notation "." in $p_{g.j}$ or $p_{gi.}$ denote summation over an index. Hui and Walter (1980) first show that this identification problem can be reduced to solving a quadratic equation and provide closed-form solutions as follows:
The false positive rates are

$$f_{Y|T^*}(1|0) = (p_{00.}p_{1.0} - p_{0.0}p_{10.} + p_{100} - p_{000} + D)/2E_0$$

$$f_{T|T^*}(1|0) = (p_{10.}p_{0.0} - p_{1.0}p_{00.} + p_{100} - p_{000} + D)/2E_1$$

The false negative rates are

$$f_{Y|T^*}(0|1) = (p_{0.1}p_{11.} - p_{01.}p_{1.1} + p_{011} - p_{111} + D)/2E_0$$

$$f_{T|T^*}(0|1) = (p_{1.1}p_{01.} - p_{11.}p_{0.1} + p_{011} - p_{111} + D)/2E_1$$

The probability of being diseased is in subpopulation $g \in \{0, 1\}$ is

$$f_{T^*|V}(1|g) = \frac{1}{2} + \{p_{g0.}(p_{0.0} - p_{1.0}) + p_{g.0}(p_{00.} - p_{10.}) + p_{100} - p_{000}\}/2D$$

where

$$E_0 = p_{1.0} - p_{0.0}$$

$$E_1 = p_{10.} - p_{00.}$$

$$D = \pm\{(p_{00.}p_{1.0} - p_{10.}p_{0.0} + p_{000} - p_{100})^2 - 4(p_{00.} - p_{10.})(p_{000}p_{1.0} - p_{100}p_{0.0})\}^{1/2}.$$

The sign of D is not determined because they don't impose the ordering assumption summarized in Hu (2008).

7.3 Measurement Errors in Quantile Regressions

7.3.1 Quantile Regressions

For a random variable Y with a monotonic CDF, its medium m_y minimizes an expected absolute deviation criterion, i.e.,

$$\begin{aligned} m_y &= \arg\min_{\beta} E[\,|Y-\beta|\,]. \\ &= \arg\min_{\beta} E[\,sign(Y-\beta)\times(Y-\beta)\,], \end{aligned} \tag{7.38}$$

where the *sign* function is defined as

$$sign(u) = 2\times(\,0.5 - \mathbf{1}(u<0)\,).$$

In this case, the first-order condition for the minimization is

$$E[\,sign(Y-m_y)\times(-1)\,] = 0, \tag{7.39}$$

which implies that $Pr(Y<m_y)=0.5$.

In a simple medium regression model, we assume

$$Y_i = X_i\beta_0 + \epsilon_i \tag{7.40}$$

where ϵ has a unique zero medium conditional on X, i.e., $E[\,\mathbf{1}(\epsilon_i<0)\,|\,X_i] = 0.5$ or equivalently $E[\,sign(\epsilon_i)\,|\,X_i] = 0$. The true parameter then satisfies

$$\beta_0 = \arg\min_{\beta} E[\,|Y_i - X_i\beta|\,], \tag{7.41}$$

with the first-order condition

$$E[\,sign(Y_i - X_i\beta_0)\times(-X_i)\,] = 0, \tag{7.42}$$

which directly implies an M-estimator.

This M-estimator can be extended to a general quantile τ by observing that the τ-th quantile of the random variable Y denoted by q_τ satisfies

$$q_\tau = \arg\min_{\beta} E[\,(\tau - \mathbf{1}(Y-\beta<0))\times(Y-\beta)\,]. \tag{7.43}$$

with the first-order condition

$$E[\,(\tau - \mathbf{1}(Y-q_\tau<0))\times(-1)\,] = 0, \tag{7.44}$$

which implies that $Pr(Y<q_\tau)=\tau$. In a simple τ-th quantile regression, we just assume that the regression error ϵ_i has a unique zero τ-th quantile conditional on X, i.e., $E[\,\mathbf{1}(\epsilon_i <$

$0) \mid X_i] = \tau$. The true parameter then satisfies

$$\beta_0 = \arg\min_{\beta} E[\,(\tau - \mathbf{1}(Y_i - X_i\beta < 0)) \times (Y_i - X_i\beta)\,], \tag{7.45}$$

with the first-order condition

$$E[\,(\tau - \mathbf{1}(Y_i - X_i\beta_0 < 0) \times (-X_i)\,] = 0, \tag{7.46}$$

which also implies an M-estimator.

7.3.2 Quantile Regressions with Measurement Errors

While there is a huge literature on measurement errors in mean regressions, the same issue has not been forgotten for quantile regressions. The relevant statistical literature tends to focus on estimation of quantile regression models when the measurement error distribution can be reasonably well estimated or belongs to a parametric family, e.g., Wei and Carroll (2009) and Wu et al. (2015). The related literature in econometrics focuses more on how both parameters of interest and the error distribution can be identified and estimated from the same data, e.g., Schennach (2008) and Firpo et al. (2017).

In this section, they provide some details in Firpo et al. (2017), which uses the Kotlarski's identity to deal with measurement errors in quantile regressions. We consider a quantile regression model as follows:

$$Y_i = X_i^*\beta_0(\tau) + Z_i\delta_0(\tau) + \epsilon_i(\tau) \tag{7.47}$$

where Y_i is the dependent variable, X_i^* is a scalar continuous covariate prone to measurement error, and Z_i is a vector of accurately-measured covariates. The τ-th quantile of the error term $\epsilon_i(\tau)$ equals zero conditional on (X_i^*, Z_i). The parameter of interest includes $(\beta_0(\tau), \delta_0(\tau))$, which satisfies

$$\begin{aligned} Q(\beta_0, \delta_0) &= E\big[\psi_\tau(Y_i - X_i^*\beta_0 - Z_i\delta_0)[X_i^*, Z_i]\big] \\ &= 0 \end{aligned} \tag{7.48}$$

with $\psi_\tau(u) = (\tau - \mathbf{1}(u < 0))$. When X_i^* is correctly observed, this equation can provide a moment condition to consistently estimate the parameter of interest. Notice that this moment condition may also be expressed as

$$\begin{aligned} Q(\beta_0, \delta_0) &= \int\int\int \psi_\tau(Y_i - X_i^*\beta_0 - Z_i\delta_0)[X_i^*, Z_i] f(X_i^*, Y_i, Z_i) dX_i^* dY_i dZ_i \\ &= E\big[\int \psi_\tau(Y_i - X^*\beta_0 - Z_i\delta_0)[X^*, Z_i] f(X^*|Y_i, Z_i) dX^*\big] \end{aligned} \tag{7.49}$$

The last expression implies that if one can identify and estimate the conditional density $f(X_i^*|Y_i, Z_i)$, then the moment condition can be used for estimation without observing X_i^*.

Firpo et al. (2017) consider the case where there are two measurements of X_i^* as follows:

$$\begin{aligned} X_{1i} &= X_i^* + U_{1i} \\ X_{2i} &= X_i^* + U_{2i}. \end{aligned} \tag{7.50}$$

The paper assumes that the two error terms U_{1i} and U_{1i} satisfies the assumptions in the Kotlarski's setting. Therefore, the distribution $f(X^*|Y,Z)$ can be identified and estimated from its characteristic function $E[e^{itX^*}|Y,Z]$ as follows:

$$E[e^{itX^*}|Y=y,Z=z] = \frac{E[e^{itX_2}|Y=y,Z=z]}{E[e^{itX_2}]} \exp\left[\int_0^t \frac{iE\left[X_1 e^{isX_2}\right]}{E\left[e^{isX_2}\right]} ds\right] \tag{7.51}$$

Notice that the right-hand side is directly estimable from the data. They then provide an estimator of (β_0, δ_0) based on the empirical moment condition:

$$\widetilde{Q}_n(\beta,\delta) = \frac{1}{n}\sum_{i=1}^n \left[\int \psi_\tau(Y_i - X^*\beta - Z_i\delta)[X^*, Z_i]\widehat{f}(X^*|Y_i,Z_i)dX^*\right]. \tag{7.52}$$

The notation $\widehat{f}(X^*|Y_i,Z_i)$ stands for the estimator of $f(X^*|Y_i,Z_i)$ specified as follows

$$\widehat{f}(X^*|Y_i,Z_i) = \frac{1}{2\pi}\int \kappa(ht)\widehat{E}[e^{itX^*}|Y=y,Z=z]exp(-itX^*)dt \tag{7.53}$$

$$\widehat{E}[e^{itX^*}|Y=y,Z=z] = \frac{\widehat{E}[e^{itX_2}|Y=y,Z=z]}{\widehat{E}[e^{itX_2}]} \exp\left[\int_0^t \frac{i\widehat{E}\left[X_1 e^{isX_2}\right]}{\widehat{E}\left[e^{isX_2}\right]} ds\right] \tag{7.54}$$

where $\widehat{E}e^{itW}$ stands for the empirical characteristic function of W, $\kappa(.)$ is the Fourier transform of a kernel function, and h is the bandwidth. The paper also shows that the estimator of (β_0, δ_0) is consistent and asymptotic normal.

8

Retrospect and Prospect

This manuscript reviews recent developments in nonparametric identification of measurement error models and their applications in microeconomic models with latent variables. The powerful identification results promote a close integration of microeconomic theory and econometric methodology, especially when latent variables are involved. With econometricians developing more application-oriented methodologies, we expect such an integration to deepen in the future research.

Besides the methodologies and the applications of measurement error models presented here, we expect this literature to advance further, with more important results. For example, the flexible nonclassical measurement error models may also provide new and convincing solutions to the endogeneity problem, a fundamental problem in econometrics. Presumably, a complete economic model should explain the causality among all the variables in the model. Endogeneity then occurs when some of the variables in the model are unobserved by the researcher. Nonclassical measurement error models may then be used to handle the unobservables, and therefore, solve the endogeneity problem under certain assumptions.

With more and more data available for researchers, we look forward to more extensive applications of the measurement error models. Given the nonparametric identification, nonparametric or semiparametric estimation of the models with latent variables may become easier than before. On the one hand, sample sizes will become much larger than before with the abundance of observations; on the other hand, researchers may observe more measurements of the latent variables. Therefore, we expect that the literature of measurement error models and their applications will keep thriving.

Bibliography

Abbring, J., P. Chiappori, and T. Zavadil, "Better Safe than Sorry? Ex Ante and Ex Post Moral Hazard in Dynamic Insurance Data," 2008. Tilburg Univeristy, working paper.

Abowd, J.M. and A. Zellner, "Estimating Gross Labor-Force Flows," *Journal of Business and Economic Statistics*, 1985, *3* (3), 254–283.

Ackerberg, D., "Advertising, Learning, and Consumer Choice in Experience Good Markets: A Structural Examination," *International Economic Review*, 2003, *44*, 1007–1040.

_ , L. Benkard, S. Berry, and A. Pakes, "Econometric Tools for Analyzing Market Outcomes," in J. Heckman and E. Leamer, eds., *Handbook of Econometrics, Vol. 6A*, North-Holland, 2007.

Adams, C., "Estimating Demand from eBay Prices," *International Journal of Industrial Organization*, 2007, *25*, 1213–1232.

Aguirregabiria, V. and P. Mira, "Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models," *Econometrica*, 2002, *70*, 1519–1543.

_ and _ , "Sequential Estimation of Dynamic Discrete Games," *Econometrica*, 2007, *75*, 1–53.

Ahn, S.C. and P. Schmidt, "Efficient Estimation of Models for Dynamic Panel Data," *Journal of Econometrics*, 1995, *68* (1), 5–28.

Ai, C. and X. Chen, "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 2003, *71* (6), 1795–1843.

Allman, Elizabeth S, Catherine Matias, and John A Rhodes, "Identifiability of Parameters in Latent Structure Models with Many Observed Variables," *The Annals of Statistics*, 2009, pp. 3099–3132.

Altonji, J.G. and R.L. Matzkin, "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 2005, *73* (4), 1053–1102.

An, Yonghong, "Identification of first-price auctions with non-equilibrium beliefs: A measurement error approach," *Journal of Econometrics*, 2017, *200* (2), 326 – 343. Measurement Error Models.

_ , Michael R Baye, Yingyao Hu, John Morgan, and Matthew Shum, "Identification and Estimation of Online Price Competition with an Unknown Number of Firms," *Journal of Applied Econometrics*, 2017, *32* (1), 80–102.

_ , **Yingyao Hu, and Matthew Shum**, "Estimating First-Price Auctions with an Unknown Number of Bidders: A Misclassification Approach," *Journal of Econometrics*, 2010, *157*, 328–341.

Anderson, T.W. and C. Hsiao, "Formulation and Estimation of Dynamic Models Using Panel Data," *Journal of Econometrics*, 1982, *18* (1), 47–82.

Andrews, D., "Examples of L^2-Complete and Boundedly-Complete Distributions," *Cowles Foundation for Research in Economics*, 2011.

Angrist, Joshua D., "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *The American Economic Review*, 1990, *80* (3), 313–336.

_ , **Guido W. Imbens, and Donald B. Rubin**, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 1996, *91* (434), 444–455.

Arcidiacono, P. and R.A. Miller, "Conditional Choice Probability Estimation of Dynamic Discrete Choice Models With Unobserved Heterogeneity," *Econometrica*, 2011, *79* (6), 1823–1867.

Arellano, M., *Panel Data Econometrics*, Oxford University Press, 2003.

_ **and B. Honore**, "Panel Data Models: Some Recent Developments," in "Handbook of Econometrics, Vol. 5," North-Holland, 2000.

_ **and O. Bover**, "Another Look at the Instrumental Variable Estimation of Error Component Models," *Journal of Econometrics*, 1995, *68* (1), 29–51.

_ **and S. Bond**, "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *The Review of Economic Studies*, 1991, *58* (2), 277–297.

Arellano, Manuel and Stéphane Bonhomme, "Robust Priors in Nonlinear Panel Data Models," *Econometrica*, 2009, *77* (2), 489–536.

_ , **Richard Blundell, and Stéphane Bonhomme**, "Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework," *Econometrica*, 2017, *85* (3), 693–734.

Athey, S. and P. Haile, "Identification of Standard Auction Models," *Econometrica*, 2002, *70*, 2107–2140.

_ , **J. Levin, and E. Seira**, "Comparing Open and Sealed Bid Auctions: Theory and Evidence from Timber Auctions," 2005. working paper, Harvard University.

Bajari, P., L. Benkard, and J. Levin, "Estimating Dynamic Models of Imperfect Competition," *Econometrica*, 2007, *75*, 1331–1370.

_ , **V. Chernozhukov, H. Hong, and D. Nekipelov**, "Nonparametric and Semiparametric Analysis of a Dynamic Game Model," 2007. Manuscript, University of Minnesota.

Banks, J. and R. Sundarum, "Denumerable-Armed Bandits," *Econometrica*, 1992, *60.*

Bassi, F. and U. Trivellato, "A latent class approach for estimating gross flows in the presence of correlated classification errors," in P. Lynn, ed., *Methodology of Longitudinal Studies*, Chichester, Wiley, 2008.

Biemer, P.P. and G. Forsman, "On the quality of reinterview data with application to the Current Population Survey," *Journal of the American Statistical Association*, 1992, *87* (420), 915–923.

_ **and J.M. Bushery**, "On the validity of Markov Latent Class Analysis for estimating classification error in labor force data," *Survey Methodology*, 2000, *26* (2), 139–152.

BLS, *Current Population Survey: Design and Methodology* Bureau of Labor Statistics 2000. Technical Paper 63RV.

Blundell, R., X. Chen, and D. Kristensen, "Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 2007, *75*, 1613–1669.

_ **,** _ **, and** _ , "Semi-nonparametric IV estimation of shape-invariant Engel curves," *Econometrica*, 2007, *75* (6), 1613–1669.

Bollinger, Christopher, "Bounding Mean Regressions When a Binary Regressor is Mismeasured," *Journal of Econometrics*, 1996, *73*, 387–399.

Bollinger, Christopher R., "Measurement Error in the Current Population Survey: A Nonparametric Look," *Journal of Labor Economics*, 1998, *16* (3), 576–594.

Bonhomme, Stéphane, Koen Jochmans, and Jean-Marc Robin, "Estimating Multivariate Latent-Structure Models," *Annals of Statistics*, 2016, *44* (2), 540–563.

_ **,** _ **, and** _ , "Non-parametric Estimation of Finite Mixtures from Repeated Measurements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016, *78* (1), 211–229.

Bordes, Laurent, Stéphane Mottelet, Pierre Vandekerkhove et al., "Semiparametric estimation of a two-component mixture model," *The Annals of Statistics*, 2006, *34* (3), 1204–1232.

Bouissou, M., J. J. Laffont, and Q. Vuong, "Tests of Noncausality under Markov Assumptions for Qualitative Panel Data," *Econometrica*, 1986, *54*, 395–414.

Bound, J., C. Brown, and N. Mathiowetz, "Measurement error in survey data," in "Handbook of Econometrics," Vol. 5 2001, pp. 3705–3843.

Bound, John, Charles Brown, and Nancy Mathiowetz, "Measurement Error in Survey Data," *Handbook of econometrics*, 2001, *5*, 3705–3843.

Buchinsky, M., J. Hahn, and J. Hotz, "Estimating Dynamic Discrete Choice Models with Heterogeneous Agents: a Cluster Analysis Approach," 2004. Working Paper, UCLA.

Canay, I.A., A. Santos, and A.M. Shaikh, "On the Testability of Identification in Some Nonparametric Models with Endogeneity," *Econometrica*, 2013, *81* (6), 2535–2559.

Canay, Ivan A., Andres Santos, and Azeem M. Shaikh, "On the Testability of Identification in Some Nonparametric Models With Endogeneity," *Econometrica*, 2013, *81* (6), 2535–2559.

Carroll, Raymond J, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective*, CRC press, 2012.

_ **, Xiaohong Chen, and Yingyao Hu**, "Identification and Estimation of Nonlinear Models Using Two Samples with Nonclassical Measurement Errors," *Journal of nonparametric statistics*, 2010, *22* (4), 379–399.

Chamberlain, G., "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 1980, *47*, 225–238.

_ , "Panel Data," *Handbook of Econometrics*, 1984, *II*, 1247–1318.

Chan, Tat Y. and Barton H. Hamilton, "Learning, Private Information, and the Economic Evaluation of Randomized Experiments," *Journal of Political Economy*, 2006, *114*, 997–1040.

Chay, K.Y., H. Hoynes, and D. Hyslop, "A Non-experimental Analysis of True State Dependence in Monthly Welfare Participation Sequences," 2001. University of California, Berkeley.

Chen, X. and D. Pouzo, "Sieve Wald and QLR Inferences On Semi/Nonparametric Conditional Moment Models," *Econometrica*, 2015, *83*, 1013–1079.

_ **and Z. Liao**, "Sieve M Inference on Irregular Parameters," *Journal of Econometrics*, 2014, *182*, 70–86.

_ **, V. Chernozhukov, S. Lee, and W. K. Newey**, "Local Identification of Nonparametric and Semiparametric Models," 2013, *Working Paper*.

Chen, Xiaohong, "Large Sample Sieve Estimation of Semi-nonparametric Models. The Handbook of Econometrics, JJ Heckman and EE Leamer (eds.), 6B," 2007.

_ **and Xiaotong Shen**, "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 1998, pp. 289–314.

_ **, H. Hong, and A. Tarozzi**, "Semiparametric efficiency in GMM models of nonclassical measurement errors, missing data and treatment effects," *Cowles Foundation Discussion Paper No. 1644, Mar. 2008*, 2008.

_ **, Han Hong, and Denis Nekipelov**, "Nonlinear Models of Measurement Errors," *Journal of Economic Literature*, 2011, *49* (4), 901–937.

_ **,** _ **, and Elie Tamer**, "Measurement Error Models with Auxiliary Data," *The Review of Economic Studies*, 2005, *72* (2), 343–366.

_ **, O. Linton, and I. Van Keilegom**, "Estimation of Semiparametric Models When the Criterion Function is Not Smooth," *Econometrica*, 2003, *71* (5), 1591–1608.

_ **, Yingyao Hu, and Arthur Lewbel**, "A Note on the Closed-form Identification of Regression Models with a Mismeasured Binary Regressor," *Statistics and Probability Letters*, 2008, *78*, 1473–1479.

_ **,** _ **, and** _ , "Nonparametric Identification and Estimation of Nonclassical Errors-in-variables Models without Additional Information," *Statistica Sinica*, 2009, *19*, 949–968.

Chernozhukov, Victor, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey, "Identification and estimation of marginal effects in nonlinear panel models," cemmap working paper CWP05/09, London 2009.

Ching, A, T Erdem, and M Keane, "Learning Models: An Assessment of Progress, Challenges, and New Developments," *Marketing Science*, 2013, *32* (6), 913–938.

Ching, Andrew T., T. Erdem, and Michael P. Keane, "Empirical Models of Learning Dynamics: A Survey of Recent Developments," *In: Wierenga B., van der Lans R. (eds) Handbook of Marketing Decision Models. International Series in Operations Research & Management Science*, 2017, *254*.

Chintagunta, P., E. Kyriazidou, and J. Perktold, "Panel Data Analysis of Household Brand Choices," *Journal of Econometrics*, 2001, *103* (1-2), 111–153.

Choi, J.J., D. Laibson, B.C. Madrian, and A. Metrick, "Reinforcement learning and savings behavior," *The Journal of Finance*, 2009, *64* (6), 2515–2534.

Compiani, Giovanni and Yuichi Kitamura, "Using mixtures in econometric models: a brief review and some new results," *The Econometrics Journal*, 2016, *19* (3).

Contoyannis, P., A.M. Jones, and N. Rice, "Simulation-based Inference in Dynamic Panel Probit Models: An Application to Health," *Empirical Economics*, 2004, *29* (1), 49–77.

Crawford, G. and M. Shum, "Uncertainty and Learning in Pharmaceutical Demand," *Econometrica*, 2005, *73*, 1137–1174.

Cuevas, A. and A. Rodríguez-Casal, "On boundary estimation," *Advances in Applied Probability*, 2004, pp. 340–354.

Cunha, Flavio, James J Heckman, and Susanne M Schennach, "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 2010, *78* (3), 883–931.

Das, S., M. Roberts, and J. Tybout, "Market Entry Costs, Producer Heterogeneity, and Export Dynamics," *Econometrica*, 2007, *75*, 837–874.

Dehejia, Rajeev H. and Sadek Wahba, "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 1999, *94* (448), 1053–1062.

Delaigle, A. and I. Gijbels, "Data-driven boundary estimation in deconvolution problems," *Computational Statistics and Data Analysis*, 2006, *50* (8), 1965–1994.

_ **and** _ , "Estimation of boundary and discontinuity points in deconvolution problems," *Statistica Sinica*, 2006, *16* (3), 773.

D'Haultfoeuille, X., "On the Completeness Condition in Nonparametric Instrumental Problems," *Econometric Theory*, 2011, *1*, 1–12.

Diamond, William and Nikhil Agarwal, "Latent indices in assortative matching models," *Quantitative Economics*, 2017, *8* (3), 685–728.

Donald, S. and H. Paarsch, "Piecewise Pseudo-Maximum Likelihood Estimation in Empirical Models of Auctions," *International Economic Review*, 1993, *34*, 121–148.

Doraszelski, U. and A. Pakes, "A Framework for Dynamic Analysis in IO," in M. Armstrong and R. Porter, eds., *Handbook of Industrial Organization, Vol. 3*, North-Holland, 2007, chapter 30.

Dunford, N. and J. Schwartz, *Linear Operators*, Vol. 3, Wiley, 1971.

Efromovich, S., *Nonparametric Curve Estimation: Methods, Theory, and Applications*, Springer-Verlag, 1999.

Erdem, T. and M. Keane, "Decision-making Under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets," *Marketing Science*, 1996, *15*, 1–20.

_ **, S. Imai, and M. Keane**, "Brand and Quantity Choice Dynamics under Price Uncertainty," *Quantitative Marketing and Economics*, 2003, *1*, 5–64.

Evdokimov, Kirill, "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity," 2010. Working paper, Princeton University.

Everitt, Brian S and J Hand David, "Finite mixture distributions," in "Monogr. Appl. Probab. Stat.," Chapman Hall, 1981.

Fan, J and Q. Yao, *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer, 2005.

Farber, H.S., "Alternative and Part-Time Employment Arrangements as a Response to Job Loss," *Journal of Labor Economics*, 1999, *17* (S4), 142–169.

Feng, Shuaizhang, "The Longitudinal Matching of Current Population Surveys: A Proposed Algorithm," *Journal of Economic and Social Measurement*, 2001, *27* (1-2), 71–91.

_ , "Longitudinal Matching of Recent Current Population Surveys: Methods, Non-matches and Mismatches," *Journal of Economic and Social Measurement*, 2008, *33* (4), 241–252.

_ **and Yingyao Hu**, "Misclassification Errors and the Underestimation of the US Unemployment Rates," *The American Economic Review*, 2013, *103* (2), 1054–1070.

Fernandez-Villaverde, J. and J. Rubio-Ramirez, "Estimating Macroeconomic Models: A Likelihood Approach," 2007. University of Pennsylvania, working paper.

Firpo, Sergio, Antonio F. Galvao, and Suyong Song, "Measurement errors in quantile regression models," *Journal of Econometrics*, 2017, *198* (1), 146 – 164.

Freyberger, Joachim, "On Completeness and Consistency in Nonparametric Instrumental Variable Models," *Econometrica*, 2017, *85* (5), 1629–1644.

_ , "Nonparametric Panel Data Models with Interactive Fixed Effects," *The Review of Economic Studies*, 2018, *85* (3), 1824–1851.

Fuller, Wayne A, *Measurement Error Models*, Vol. 305, John Wiley & Sons, 2009.

Ghahramani, Z., "An Introduction to Hidden Markov Models and Bayesian Networks," *International Journal of Pattern Recognition and Artificial Intelligence*, 2001, *15*, 9–42.

Gittins, J. and G. Jones, "A Dynamic Allocation Index for the Sequential Design of Experiments," in et. al. J. Gani, ed., *Progress in Statistics*, North-Holland, 1974.

Gourieroux, C. and A. Monfort, "Simulation-based Inference: A Survey with Special Reference to Panel Data Models," *Journal of Econometrics*, 1993, *59* (1-2), 5–33.

Guerre, Emmanuel, Isabelle Perrigne, and Quang Vuong, "Optimal Nonparametric Estimation of First-price Auctions," *Econometrica*, 2000, *68* (3), 525–574.

_ , _ , and _ , "Nonparametric Identification of Risk Aversion in First-price Auctions under Exclusion Restrictions," *Econometrica*, 2009, *77* (4), 1193–1227.

Hahn, Jinyong, "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 1998, *66* (2), 315–331.

_ , "The Information Bound of a Dynamic Panel Logit Model with Fixed Effects," *Econometric Theory*, 2001, *17* (5), 913–932.

_ **and Zhipeng Liao**, "Nonparametric Instrumental Variables And Regular Estimation," *Econometric Theory*, 2018, *34* (3), 574–597.

Haile, P., H. Hong, and M. Shum, "Nonparametric Tests for Common Values in First-Price Auctions," 2003. NBER working paper #10105.

Hajivassiliou, V. and P. Ruud, "Classical Estimation Methods for LDV Models Using Simulation," *Handbook of econometrics*, 1994, *4*, 2383–2441.

Hajivassiliou, V.A., "Simulation Estimation Methods for Limited Dependent Variable Models," *Handbook of Statistics*, 1993, *59.*

Hall, P. and X.H. Zhou, "Nonparametric estimation of component distributions in a multivariate mixture," *ANNALS OF STATISTICS*, 2003, *31* (1), 201–224.

_ , I. McKay, and B. Turlach, "Performance of Wavelet Methods for Functions with Many Discontinuities," *Annals of Statistics*, 1996, *24*, 2462–2476.

Halliday, T., "Heterogeneity, State Dependence and Health," 2002. Princeton University.

Hampton, A., P. Bossaerts, and J. O'Doherty, "The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans," *Journal of Neuroscience*, 2006, *26*, 8360–8367.

Heckman, J.J., "Simple Statistical Models for Discrete Panel Data Developed and Applied to Test the Hypothesis of True State Dependence Against the Hypothesis of Spurious State Dependence," in "Annales de l'INSEE" Institut national de la statistique et des études économiques 1978, pp. 227–269.

_ , "Statistical Models for Discrete Panel Data," *in Structural Analysis of Discrete Panel Data with Econometric Applications, ed. by C. Manski and D. McFadden. Cambridge: MIT Press*, 1981, pp. 179–195.

_ , "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process," *Structural Analysis of Discrete Data with Econometric Applications*, 1981, pp. 114–178.

_ **and E. Vytlacil**, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 2005, pp. 669–738.

_ **and R.J. Willis**, "A Beta-logistic Model for the Analysis of Sequential Labor Force Participation by Married Women," *The Journal of Political Economy*, 1977, pp. 27–58.

Hendel, I. and A. Nevo, "Measuring the Implications of Sales and Consumer Stockpiling Behavior," *Econometrica*, 2006, *74*, 1637–1674.

Hendricks, K., J. Pinkse, and R. Porter, "Empirical Implications of Equilibrium Bidding in First-Price, Symmetric, Common-Value Auctions," *Review of Economic Studies*, 2003, *70*, 115–145.

Henry, M., Y. Kitamura, and B. Salanie, "Nonparametric identification of mixtures with exclusion restrictions," 2008. University of Montreal, working paper.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 2003, *71* (4), 1161–1189.

Hoderlein, S. and E. Mammen, "Identification of Marginal Effects in Nonseparable Models without Monotonicity," *Econometrica*, 2007, *75* (5), 1513–1518.

_ **and H. White**, "Nonparametric Identification in Nonseparable Panel Data Models With Generalized Fixed Effects," *CeMMAP Working Paper*, 2009.

Hong, H. and M. Shum, "Increasing Competition and the Winner's Curse: Evidence from Procurement," *Review of Economic Studies*, 2002, *69*, 871–898.

_ **and** _ , "Pairwise-Difference Estimation of a Dynamic Optimization Model," 2007. Revised manuscript, Stanford University.

Honoré, B.E., "Orthogonality Conditions for Tobit Models with Fixed Effects and Lagged Dependent Variables," *Journal of Econometrics*, 1993, *59* (1-2), 35–61.

Honoré, B.E. and E. Kyriazidou, "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 2000, *68* (4), 839–874.

_ **and E. Tamer**, "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 2006, *74* (3), 611–629.

_ **and L. Hu**, "Estimation of Cross Sectional and Panel Data Censored Regression Models with Endogeneity," *Journal of Econometrics*, 2004, *122* (2), 293–316.

Hotz, J., R. Miller, S. Sanders, and J. Smith, "A Simulation Estimator for Dynamic Models of Discrete Choice," *Review of Economic Studies*, 1994, *61*, 265–289.

Hotz, V. Joseph and Robert A. Miller, "Conditional Choice Probabilties and the Estimation of Dynamic Models," *Review of Economic Studies*, 1993, *60*, 497–529.

Houde, J.-F. and S. Imai, "Identification and 2-step Estimation of DDC models with Unobserved Heterogeneity," 2006. Working Paper, Queen's University.

Hu, L., "Estimation of a Censored Dynamic Panel Data Model," *Econometrica*, 2002, *70* (6), 2499–2517.

Hu, Yingyao, "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution," *Journal of Econometrics*, 2008, *144*, 27–61.

_ , "The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics," *Journal of Econometrics*, 2017, *200*, 154–168.

_ **and Ji-Liang Shiu**, "Nonparametric Identification Using Instrumental Variables: Sufficient Conditions For Completeness," *Econometric Theory*, 2018, *34* (3), 659–693.

_ **and Matthew Shum**, "Nonparametric Identification of Dynamic Models with Unobserved State Variables, Supplemental Material," 2009. available at *http://www.hss.caltech.edu/ ~mshum/papers/onlineapp.pdf*.

_ **and** _ , "Nonparametric Identification of Dynamic Models with Unobserved State Variables," *Journal of Econometrics*, 2012, *171* (1), 32–44.

_ **and** _ , "Identifying Dynamic Games with Serially-correlated Unobservables," *Advances in Econometrics*, 2013, *31*, 97–113.

_ **and Ruli Xiao**, "Global estimation of finite mixture and misclassification models with an application to multiple equilibria," *Econometric Review*, forthcoming.

_ **and Susanne Schennach**, "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 2008, *76*, 195–216.

_ **and Yuya Sasaki**, "Closed-Form Estimation of Nonparametric Models with Non-Classical Measurement Errors," *Journal of Econometrics*, 2015, *185* (2), 392–408.

_ **and** _ , "Identification of Paired Nonseparable Measurement Error Models," *Econometric Theory*, 2017, *33* (4), 955–979.

_ **and** _ , "Closed-Form Identification of Dynamic Discrete Choice Models with Proxies for Unobserved State Variables," *Econometric Theory*, 2018, *34* (1), 166–185.

_ **, David McAdams, and Matthew Shum**, "Nonparametric Identification of First-price Auction Models with Non-Separable Unobserved Heterogeneity," *Journal of Econometrics*, 2013, *174*, 186–193.

_ **, Robert Moffitt, and Yuya Sasaki**, "Semiparametric Estimation of the Canonical Permanent-Transitory Model of Earnings Dynamics," *Quantitative Economics*, 2018.

_ **, Susanne Schennach, and Ji-Liang Shiu**, "Injectivity of a Class of Integral Operators with Compactly Supported Kernels," *Journal of Econometrics, forthcoming*, 2017.

_ , _ , **and** _ , "Identification of nonparametric monotonic regression models with continuous nonclassical measurement errors," *Journal of Econometrics, forthcoming*, 2021.

_ , **Yutaka Kayaba, and Matthew Shum**, "Nonparametric Learning Rules from Bandit Experiments: The Eyes Have It!," *Games and Economic Behavior*, 2013, *81*, 215–231.

Hui, S.L. and S.D. Walter, "Estimating the Error Rates of Diagnostic Tests," *Biometrics*, 1980, *36*, 167–171.

Hunter, David R, Shaoli Wang, and Thomas P Hettmansperger, "Inference for mixtures of symmetric distributions," *The Annals of Statistics*, 2007, pp. 224–251.

Hyslop, D.R., "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Participation Behavior: Monte Carlo Evidence and Empirical Results for Married Women," *Industrial Relations Section Working Paper #347, Princeton University*, 1995.

_ , "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women," *Econometrica*, 1999, *67* (6), 1255–1294.

Imai, S., N. Jain, and A. Ching, "Bayesian estimation of dynamic discrete choice models," *Econometrica*, 2009, *77* (6), 1865–1899.

Jones, S.R.G. and C.W. Riddell, "The measurement of unemployment: an empirical approach," *Econometrica*, 1999, *67* (1), 147–162.

Kane, Thomas, Cecilia Rouse, and Douglas Staiger, "Estimating returns to schooling when schooling is misreported," 1999. Working paper no. 7235, NBER.

Kasahara, Hiroyuki and Katsumi Shimotsu, "Nonparametric identification of finite mixture models of dynamic discrete choices," *Econometrica*, 2009, *77* (1), 135–175.

Keane, Michael P and Kenneth I Wolpin, "The career decisions of young men," *Journal of political Economy*, 1997, *105* (3), 473–522.

Keane, M.P., "Simulation Estimation for Panel Data Models with Limited Dependent Variables," *Handbook of Statistics*, 1993, *11*, 545–571.

Kotlarski, Ignacy, "On Pairs of Independent Random Variables Whose Product Follows the Gamma Distribution," *Biometrika*, 1965, pp. 289–294.

Krasnokutskaya, E. and K. Seim, "Bid Preference Programs and Participation in Highway Procurement Auctions," *American Economic Review*, 2011, *101* (6), 2653–86.

Krasnokutskaya, Elena, "Identification and Estimation in Procurement Auctions under Unobserved Auction Heterogeneity," *Review of Economic Studies*, 2011, *28*, 293–327.

Laffont, J. J., H. Ossard, and Q. Vuong, "Econometrics of First-Price Auctions," *Econometrica*, 1995, *63*, 953–980.

LaLonde, Robert J., "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review*, 1986, *76* (4), 604–620.

Levinsohn, J. and A. Petrin, "Estimating Production Functions Using Intermediate Inputs to Control for Unobservables," 2000. Manuscript, University of Michigan.

Lewbel, Arthur, "Estimation of Average Treatment Effects with Misclassification," *Econometrica*, 2007, *75* (2), 537–551.

Li, Tong, "Robust and consistent estimation of nonlinear errors-in-variables models," *Journal of Econometrics*, 2002, *110* (1), 1–26.

_ , "Econometrics of First Price Auctions with Entry and Binding Reservation Prices," *Journal of Econometrics*, 2005, *126*, 173–200.

_ **and Quang Vuong**, "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 1998, *65* (2), 139–165.

_ **and X. Zheng**, "Entry and Competition Effects in First-Price Auctions: Theory and evidence from Procurement Auctions," 2006. working paper, Vanderbilt University.

_ **, I. Perrigne, and Q. Vuong**, "Structural Estimation of the Affiliated Private Value Acution Model," *RAND Journal of Economics*, 2002, *33*, 171–193.

_ **, Isabelle Perrigne, and Quang Vuong**, "Conditionally Independent Private Information in OCS Wildcat Auctions," *Journal of Econometrics*, 2000, *98*, 129–161.

Madrian, B. and L. Lefgren, "An Approach to Longitudinally matching Current Population Survey (CPS) Respondents," *Journal of Economic and Social Measurement*, 2000, *26*, 31–62.

Magnac, T. and D. Thesmar, "Identifying Dynamic Discrete Decision Processes," *Econometrica*, 2002, *70*, 801–816.

_ **and M. Visser**, "Transition Models with measurement errors," *Review of Economics and Statistics*, 1999, *81* (3), 466–474.

Mahajan, Aprajit, "Identification and Estimation of Regression Models with Misclassification," *Econometrica*, 2006, pp. 631–665.

Marcoul, Philippe and Quinn Weninger, "Search and active learning with correlated information: Empirical evidence from mid-Atlantic clam fishermen," *Journal of Economic Dynamics and Control*, 2008, *32*, 1921–1948.

Marmer, V., A. Shneyerov, and P. Xu, "What Model for Entry in First-Price Auctions? A Nonparametric Approach," *Journal of Econometrics*, 2013, *176* (1), 46–58.

Mattner, L., "Some Incomplete but Boundedly Complete Location Families," *The Annals of Statistics*, 1993, *21*, 2158–2162.

Matzkin, R., "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 2003, *71*, 1339–1376.

Meyer, B.D., "Classification-error models and labor-market dynamics," *Journal of Business and Economic Statistics*, 1988, *6* (3), 385–390.

Miller, R., "Job Matching and Occupational Choice," *Journal of Political Economy*, 1984, *92*, 1086–1120.

Molinari, Francesca, "Partial Identification of Probability Distributions with Misclassified Data," *Journal of Econometrics*, 2008, *144*, 81–117.

Newey, W.K. and J.L. Powell, "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 2003, *71* (5), 1565–1578.

Neyman, J. and E.L. Scott, "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 1948, *93*, 1–32.

Norets, A., "Inference in dynamic discrete choice models with serially correlated unobserved state variables," *Econometrica*, 2009, *77*, 1665–1682.

Odean, T., M. Strahilevitz, and B. Barber, "Once Burned, Twice Shy: How Naive Learning and Counterfactuals Affect the Repurchase of Stocks Previously Sold," Technical Report 2004. mimeo., UC Berkeley, Haas School.

Olley, S. and A. Pakes, "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 1996, *64*, 1263–1297.

Paarsch, H., "Deriving an Estimate of the Optimal Reserve Price: An Application to British Columbian Timber Sales," *Journal of Econometrics*, 1997, *78*, 333–357.

_ **and H. Hong**, *An Introduction to the Structural Econometrics of Auction Data*, MIT Press, 2006. with M. Haley.

Pagan, A. and A. Ullah, *Nonparametric Econometrics*, Cambridge University Press, 1999.

Pakes, A., "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, 1986, *54* (4), 755–84.

_ **and M. Simpson**, "Patent Renewal Data," *Brookings Papers on Economic Activity: Microeconomics*, 1989, pp. 331–403.

_ **, M. Ostrovsky, and S. Berry**, "Simple Estimators for the Parameters of Discrete Dynamic Games (with Entry/Exit Examples)," *RAND Journal of Economics*, 2007, *37*.

Peracchi, F. and F. Welch, "How Representative Are Matched Cross Sections? Evidence from the Current Population Survey," *Journal of Econometrics*, 1995, *68* (1), 153–179.

Pesendorfer, M. and P. Schmidt-Dengler, "Asymptotic Least Squares Estimators for Dynamic Games," *Review of Economic Studies*, 2008, *75* (3), 901–928.

Poterba, J.M. and L.H. Summers, "Response variation in the CPS: caveats for unemployment analysts," *Monthly Labor Review*, March 1984, pp. 37–43.

_ **and** _ , "Reporting Errors and Labor Market Dynamics," *Econometrica*, 1986, *54* (6), 1319–1338.

_ **and** _ , "Unemployment benefits and labor market transitions: a multinomial logit model with errors in classification," *Review of Economics and Statistics*, 1995, *77* (2), 207–216.

Rao, B.L.S. Prakasa, *Identifiability in Stochastic Models: Characterization of Probability Distributions*, Academic Press, Inc., 1992.

Reiersøl, Olav, "Identifiability of a Linear Relation Between Variables Which are Subject to Error," *Econometrica*, 1950, pp. 375–389.

Rosenbaum, Paul R. and Donald B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 04 1983, *70* (1), 41–55.

Rust, J., "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica*, 1987, *55*, 999–1033.

_ , "Structural Estimation of Markov Decision Processes," in R. Engle and D. McFadden, eds., *Handbook of Econometrics, Vol. 4*, North Holland, 1994, pp. 3082–146.

Samuelson, W. F., "Competitive Bidding with Entry Costs," *Economics Letters*, 1985, *17*, 53–57.

Sasaki, Yuya, "Heterogeneity and Selection in Dynamic Panel Data," *Journal of Econometrics*, 2015, *188*, 236–249.

Schennach, S.M., "Quantile regression with mismeasured covariates," *Econometric Theory*, 2008, *24* (4), 1010–1043. cited By 23.

Schennach, Susanne M, "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 2004, *72* (1), 33–75.

_ , "Nonparametric Regression in the Presence of Measurement Error," *Econometric Theory*, 2004, *20* (06), 1046–1093.

_ , "Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models," *Econometrica*, 2007, *75* (1), 201–239.

_ , "Regressions with Berkson Errors in Covariates: A Nonparametric Approach," *The Annals of Statistics*, 2013, *41* (3), 1642–1668.

_ , "Recent Advances in the Measurement Error Literature," *Annual Review of Economics*, 2016, *8*, 341–377.

_ , "Mismeasured and Unobserved Variables," *Handbook of Econometrics*, 2019.

_ **and Yingyao Hu**, "Nonparametric Identification and Semiparametric Estimation of Classical Measurement Error Models without Side Information," *Journal of the American Statistical Association*, 2013, *108* (501), 177–186.

Shen, X. and W.H. Wong, "Convergence Rate of Sieve Estimates," *Annals of Statistics*, 1994, *22*, 580–615.

Shen, Xiaotong, "On Methods of Sieves and Penalization," *The Annals of Statistics*, 1997, *25* (6), 2555–2591.

Shiu, Ji-Liang and Yingyao Hu, "Identification and Estimation of Nonlinear Dynamic Panel Data Models with Unobserved Covariates," *Journal of Econometrics*, 2013, *175* (2), 116–131.

Sinclair, M.D. and J.L. Gastwirth, "On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data," *Journal of the American Statistical Association*, 1996, *91* (435), 961–969.

_ **and** _ , "Estimates of the errors in classification in the labour force survey and their effect on the reported unemployment rate," *Survey methodology*, 1998, *24* (2), 157–169.

Singh, A.C. and J.N.K. Rao, "On the adjustment of gross flow estimates for classification error with application to data from the Canadian labour force survey," *Journal of the American Statistical Association*, 1995, *90* (430), 478–488.

Song, U., "Nonparametric Estimation of an E-Bay Auction Model with an Unknown Number of Bidders," 2004. working paper, University of British Columbia.

_ , "Nonparametric Identification and Estimation of a First-Price Auction Model with an Uncertain Number of Bidders," 2006. working paper, University of British Columbia.

Stahl, Dale O. and Paul W. Wilson, "Experimental evidence on players' models of other players," *Journal of Economic Behavior & Organization*, 1994, *25* (3), 309 – 327.

Sutton, R. and A. Barto, *Reinforcement Learning*, MIT Press, 1998.

van den Berg, G. and B. van der Klaauw, "If Winning Isn't Everything, Why do they Keep Score? A Structural Empirical Analysis of Dutch Flower Auctions," 2007. mimeo, Free University Amsterdam.

Wansbeek, Thomas J. and Erik Meijer, *Measurement Error and Latent Variables in Econometrics* Advanced textbooks in economics, Elsevier, 2000.

Wei, Ying and Raymond J. Carroll, "Quantile Regression With Measurement Error," *Journal of the American Statistical Association*, 2009, *104* (487), 1129–1143.

Wilhelm, Daniel, "Identification and Estimation of Nonparametric Panel Data Regressions with Measurement Error," Technical Report, University College London 2013.

Wooldridge, J.M., "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," *Journal of Applied Econometrics*, 2005, *20* (1), 39–54.

_ , *Econometric Analysis of Cross Section and Panel Data*, The MIT press, 2010.

Wu, Yuanshan, Yanyuan Ma, and Guosheng Yin, "Smoothed and Corrected Score Approach to Censored Quantile Regression With Measurement Errors," *Journal of the American Statistical Association*, 2015, *110* (512), 1670–1683.

Xiao, Ruli, "Identification and Estimation of Incomplete Information Games with Multiple Equilibria," *Journal of Econometrics*, 2018, *203* (2), 328–343.

Xin, Yi, "Asymmetric Information, Reputation, and Welfare in Online Credit Markets," *Working paper, Caltech*, 2018.

Xu, D., "A Structural Empirical Model of R&D, Firm Heterogeneity, and Industry Evolution," 2007. Manuscript, New York University.